Empirical Investigation of Randomized Quantile Residuals for Diagnosis of Non-Normal Regression Models

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon

By

Alireza Sadeghpour

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Mathematics and Statistics
142 Mcclean Hall
106 Wiggins Road
University of Saskatchewan
Saskatoon, SK Canada
S7N 5E6

Abstract

Traditional tools for model diagnosis for Generalized Linear Model (GLM), such as deviance and Pearson residuals, have been often utilized to examine goodness of fit of GLMs. In normal linear regression, both of these residuals coincide and are normally distributed; however in non-normal regression models, such as Logistic or Poisson regressions, the residuals are far from normality, with residuals aligning nearly parallel curves according to distinct response values, which imposes great challenges for visual inspection. As such, the residual plots for modeling discrete outcome variables convey very limited meaningful information, which render it of limited practical use.

Randomized quantile residuals was proposed in literature to circumvent the above-mentioned problems in the traditional residuals in modeling discrete outcomes. However, this approach has not gained deserved awareness and attention. Therefore, in this thesis, we theoretically justify the normality of the randomized quantile residuals and compare their performance with the traditional ones, Pearson and deviance residuals, through a set of simulation studies. Our simulation studies demonstrate the normality of randomized quantile residuals when the fitted model is true. Further, we show that randomized quantile residual is able to detect many kinds of model inadequacies. For instance, the linearity assumption of the covariate effect in GLM can be examined by visually checking the plots of randomized quantile residuals against the predicted values or the covariates. Randomized quantile residuals can be also used to detect overdispersion and zero-inflation, two commonly occurred cases associated with count data. We advocate examining normality of the randomized quantile residuals as a unifying way for examining the goodness of fit for regression model, especially for modeling the discrete outcomes. We also demonstrate this approach in a real application studying the independent association between air pollution and daily influenza incidence in Beijing, China.

Acknowledgements

I would like to first and foremost express my sincere gratitude to my supervisors, Dr. Longhai Li and Dr. Cindy Feng, for their academic and financial support as well as their encouragements and patience throughout the period of my MSc. education. It was an honour to work with two of the most dedicated academics I have ever met.

I like to thank Dr. Chris Soteros for agreeing to serve as my thesis supervisor when I applied to study for my MSc. in Canada. It was her willingness to trust in my academic abilities which allowed me to have the opportunity to study at the University of Saskatchewan in Canada. I like to thank Dr. Shahedul Khan for serving as a member of my thesis committee as well as teaching me various statistical skills which will be useful in my career as a statistician.

My special thanks goes to my dear friend Mehdi Rostami for all the wonderful friendship and companionship we have had in the last nine years. His advice and guidance in Mathematics, Statistics, and computer programming have always been very insightful in all aspects of my academic life.

I would like to thank the Department of Mathematics and Statistics at the University of Saskatchewan for academic and financial support as well as opportunities given to me. It has been a wonderful two years.

Finally, I would specially like to express my very profound gratitude to my family for providing me with continuous love and support during my whole life. Their love, enthusiasm, and encouragements have always inspired me to strive towards my goals. My accomplishments would not have been possible without them. Thank you for everything.

Dedicated to my father, Mohammadreza Sadeghpour

Contents

List of Tables

List of Figures

List of Abbreviations

GLM	Generalized Linear Model
ZIP	Zero-Inflated Poisson
GOF	Goodness of Fit
CDF	Cumulative Distribution Function
PDF	Probability Density Function
PMF	Probability Mass Function
LOOCV	Leave-One-Out Cross-Calidation
PM	Particulate Matter
ILI	Influenza-Like-Illnesses
GAM	Generalized Additive Model
AIC	Akaikes Information Criterion

Chapter 1

Introduction

Examining residuals is a primary method to identify the discrepancies between models and data. Deviance and Pearson residuals have been often used for model diagnosis for generalized linear models (GLM). In normal linear regression, these residuals coincide and are normally distributed; however in modeling discrete outcome variables, for example count data, the residuals are far from normality, forming nearly parallel curves, leading to difficulty for visual inspection and interpretation. As such, the residual plots for modeling discrete outcome variables are not informative and may be misleading.

Randomized quantile residuals, defined by Dunn and Smyth in 1996 [15], remedy the above-mentioned problems of the traditional residuals for modeling discrete outcome variable. However, only a few researchers have devoted their attention to the quantile residuals [7, 14, 31, 38] and their potential for examining goodness of fit for a wide range of models have never been studied in detail or completely exploited. In this thesis, we will theoretically prove that the randomized quantile residuals are normally distributed. We will also demonstrate their superior performance over the Pearson and deviance residuals through simulation studies. We will also apply quantile residuals and the traditional residuals to a real application.

1.1 Review of Traditional Residuals

The GLM framework [30] generalizes the ordinary linear regression allowing the response variable following non-normal distribution, such as Poisson, Gamma, negative binomial, and etc. All these distributions belong to a broad family, called *exponential dispersion family*, having the probability density function (PDF) or probability mass function (PMF) of the

form:

$$f(y_i;\theta_i,\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\} \tag{1.1}$$

for some functions a, b, and c. More details will be given in Chapter 2. In GLM, a link function is used to connect the expected value of the response variable to a linear combination of the covariates and regression parameters [1] as

$$g(E(\mathbf{y})) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \tag{1.2}$$

where $E(\mathbf{y})$ is the expected value of $\mathbf{y}$, a vector of y_i, $i = 1, \cdots, n$. $\mathbf{X}$ is the model matrix that contains the value of all explanatory variables, and $\boldsymbol{\beta}$ is the parameter vector.

In GLM, two types of residuals have been used traditionally: Pearson and deviance residuals [29].

Pearson residuals measure the standardized distance between an observed and expected response directly. Even though the Pearson residual has mean and standard deviation of 0 and 1, its distribution is often skewed and is not normally distributed, which makes it difficult to visually decide about the model adequacy [29]. The *deviance residual* is defined as signed square root of the individual contribution to the deviance of the model, the difference between log-likelihood of fitted model and saturated model (the model with perfect fit). Pierce and Schafer (1986) [32] indicated that the deviance residuals should be more nearly normal than the Pearson, but again when data is highly dispersed relative to the mean, neither deviance nor Pearson residuals follows normal distribution. Another drawback of deviance residual is that it is sometimes challenging to define deviance residuals for some complex models. Last but not least, Pearson and deviance residuals usually have variance less than 1 because instead of comparing with the true mean μ_i, they compare y_i with the fitted mean $\hat{\mu}_i$ [1, 32]. Another disadvantage is that for count data, it often does not provide any interpretable results for model diagnostics. For example, in Poisson regression, the response variable typically takes on limited number of unique values, so the residual plot for both Pearson and deviance forms nearly parallel curves corresponding to distinct response values, which provides limited meaningful information [15].

1.2 Randomized Quantile Residuals

To circumvent the difficulty of interpreting the traditional residuals, *randomized quantile residuals* were proposed by Dunn and Smyth in 1996 [15]. The central idea is to map the discrete outcome variable by using some uniform random variable. When the outcome variable is a continuous random variable. The randomized quantile residual is defined as the probit transformation of the cumulative distribution of the response variable. In discrete case, randomization will be imposed to make the cumulative distribution function (CDF) continuous. To be more specific, let $F(Y;\mu,\phi)$ be the CDF of the random variable Y. In discrete case, let also $p(Y;\mu,\phi)$ be the PMF of Y. Suppose

$$F_i^* = \begin{cases} F(y_i;\hat{\mu}_i,\hat{\phi}_i) & F \text{ is continuous} \\ F(y_i^-;\hat{\mu}_i,\hat{\phi}_i) + u_i\, p(y_i;\hat{\mu}_i,\hat{\phi}_i) & F \text{ is not continuous} \end{cases} \tag{1.3}$$

where $\hat{\mu}_i$ and $\hat{\phi}_i$ are estmation of mean and dispersion parameter and $F(y_i^-;\hat{\mu}_i,\hat{\phi}_i) = \lim_{y\to y_i^-} F(y;\hat{\mu}_i,\hat{\phi}_i)$ and u_i is a uniform random variable on $(0,1]$. Then, the randomized quantile residual [15], q_i, is defined as

$$q_i = \Phi^{-1}(F_i^*) \tag{1.4}$$

where $\Phi()$ is the CDF of the standard normal distribution.

We will show in Chapter 2 that q_i is standard normal given the known true parameters μ_i and ϕ_i. On the other hand, Pearson and deviance residuals are not necessarily normal when the data is highly dispersed relative to the mean and their distribution are mostly skewed [15]. In discrete case, this result is important because it enables us to visually check the residuals, while the Pearson and deviance residuals are usually uninformative.

Moreover, randomized quantile residuals can be applied to model the response variable following different types of distributions in a unified way, which makes calculation much easier than deviance residuals. The only information needed for computing randomized quantile residual is knowing the CDF of the response variable. This is a great advantage comparing deviance residuals, which might be challenging to find in more complex models.

The randomized quantile residuals can be also applied for model diagnosis when the response variable does not belong to the GLM. In this thesis, we will demonstrate this

extension for the zero-inflated models, such as zero-inflated Poisson, models that are used for modeling the response variable with a excessive mass at zero as compared to a usual count distribution [1, 22]. Similar to other count models, Pearson residuals fail to give much information in those models with a high percentage of zeros. The saturated model for zero-inflated model is not easily defined, and in case we can find deviance residuals, it is far from being normal even if the model is true. Nevertheless, we can easily compute the quantile residuals for zero-inflated models which will be shown in this thesis that are more normally distributed than the traditional residuals.

1.3 Contributions of this thesis

In the remaining of this thesis, we review different residuals for GLM and for zero-inflated models in Chapter 2. We discuss how the traditional residuals like Pearson and deviance residuals fail to provide useful information for modeling discrete data. Then, we define the randomized quantile residuals and prove theoretically that randomized quantile residuals follow a standard normal distribution under the true model, even for the discrete outcome variable. In the last Section of Chapter 2, we will review some normality tests for examining the normality of the residuals.

Our main purpose of this research is to compare the randomized quantile residuals versus the transitional residuals using both simulated and real datasets. In Chapter 3, we consider three scenarios in model diagnosis that are commonly encountered in real applications, i.e. non-linearity in covariate effect, overdispersion, and zero-inflation. We demonstrate how randomized quantile residuals can be a unified and useful tool for model diagnosis, especially for modeling discrete data as compared with the traditional residuals.

In Chapter 4, we demonstrate the advantage of the randomized quantile residuals by applying them to a real application studying the independent association between air pollution and influenza incidence in Beijing, China. Concluding remarks are given in Chapter 5.

Chapter 2

Residuals for Model Diagnostics

GLM is a unifying conceptual framework encompassing various statistical models for modeling not only normal, but also non-normal data. Another type of non-normal response variable that often occur in practice is zero-inflated model. In Sections 2.1 and 2.2, these two models will be introduced briefly. In Section 2.3, traditional residuals for model diagnosis in these models, namely Pearson and deviance will be reviewed. In Section 2.4, an introduction of randomized quantile residuals along with the theoretical proof for their normality will be provided. Finally, in Section 2.5, common normality tests will be briefly introduced for checking normality of the residuals.

2.1 GLM

GLM [1] is an extension of ordinary linear models which allows the response variable follow a non-normal distribution. GLM consists of three components:

- *Random component*
- *Linear predictor*
- *Link function*

The *random component* specifies the distribution of the response variable with independent observations $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ conditional on the explanatory variables arranged in a model matrix $\boldsymbol{X}$. In GLM, we assume the distribution with probability density or mass

function as the following form

$$f(y_i;\theta_i,\phi)=\exp\left\{\frac{y_i\theta_i-b(\theta_i)}{a(\phi)}+c(y_i,\phi)\right\} \tag{2.1}$$

Any density of the above form is called the *exponential dispersion family*. The parameters θ_i and ϕ are called the *natural parameter* and the *dispersion parameter*, respectively, and a, b, and c are arbitrary functions.

For a parameter vector $\boldsymbol{\beta}=(\beta_1,\beta_2,\ldots,\beta_p)^T$ and a $n\times p$ model matrix $\boldsymbol{X}$ of p explanatory variables, $\boldsymbol{\eta}=\mathbf{X}\boldsymbol{\beta}$ is called the *linear predictor*.

The *link function*, a monotonic differentiable function g, connects random component of a GLM with a linear combination of the covariates and regression parameters by

$$g(E(\boldsymbol{y}))=\boldsymbol{\eta}=\mathbf{X}\boldsymbol{\beta} \tag{2.2}$$

The inverse function of g, g^{-1}, is called the *response function*. If g maps the mean to the natural parameter, i.e. $g(\mu_i)=\theta_i$, then it is called the *canonical link*. In this case,

$$\theta_i=g\left(\mu_i\right)=\eta_i=\sum_{i=1}^{p}\beta_j x_{ij}. \tag{2.3}$$

One of the advantages of exponential dispersion family is that it satisfies some regularity conditions (such as differentiation passing under an integral sign), which enables us to calculate expected value and variance of random component easily as followings:

Let $L_i=\log f(y_i;\theta_i,\phi)$ denote the contribution of y_i to the log-likelihood function, $L=\sum L_i$. Then, we have

$$\begin{aligned}
L_i&=\frac{y_i\theta_i-b(\theta_i)}{a(\phi)}+c(y_i,\phi)\\
\frac{\partial L_i}{\partial\theta_i}&=\frac{y_i-b'(\theta_i)}{a(\phi)}\\
\frac{\partial^2 L_i}{\partial\theta_i^2}&=\frac{-b''(\theta_i)}{a(\phi)},
\end{aligned}$$

where $b'(\theta_i)$ and $b''(\theta_i)$ shows first and second derivative of b calculated at θ_i.

Now, because regularity conditions hold for exponential dispersion family, we have:

$$E\left(\frac{\partial L_i}{\partial\theta_i}\right)=0 \qquad \text{and} \qquad -E\left(\frac{\partial^2 L_i}{\partial\theta_i^2}\right)=E\left(\frac{\partial L_i}{\partial\theta_i}\right)^2$$

Combining the above results, we can find the mean and variance of the exponential dispersion family as followings:

$$\mu_i = E(y_i) = b'(\theta_i) \tag{2.4}$$

$$V(y_i) = b''(\theta_i)a(\phi) \tag{2.5}$$

In the following, a brief introduction of some special cases of GLM, for example Poisson, negative binomial, and Gamma will be given.

2.1.1 Poisson

Suppose y_i, $i = 1, \cdots, n$ follows a Poisson distribution, then its probability mass function is

$$f(y_i; \mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = \exp\{y_i \log \mu_i - \mu_i - \log(y_i!)\} \tag{2.6}$$

We denote $Poisson(\mu_i)$ as a Poisson distribution with parameter μ_i. Let $dpois(y_i; \mu_i)$ and $ppois(y_i; \mu_i)$ denote its PMF and CDF, respectively. Now considering the natural parameter $\theta_i = \log \mu_i$, $b(\theta_i) = \exp(\theta_i) = \mu_i$, $a(\phi) = 1$, and $c(y_i, \phi) = -\log(y_i!)$, then y_i has exponential dispersion form defined by equation (2.1). By equations (2.4) and (2.5),

$$E(y_i) = b'(\theta_i) = \exp(\theta_i) = \mu_i \tag{2.7}$$

$$V(y_i) = b''(\theta_i)a(\phi) = \exp(\theta_i) \times 1 = \mu_i \tag{2.8}$$

2.1.2 Negative Binomial

Poisson distribution assumes that the mean and the variance are the same. However, there are lots of situations where the variance is greater than the mean (overdispersion) or the variance is less than the mean (underdispersion). One possible solution to capture overdisperssion is using negative binomial instead of Poisson. Suppose y_i has a negative binomial distribution with parameters μ_i and k, then its probability mass function is

$$f(y_i; \mu_i, k) = \frac{\Gamma(y_i + k)}{\Gamma(k)\Gamma(y_i + 1)} \left(\frac{\mu_i}{\mu_i + k}\right)^{y_i} \left(\frac{k}{\mu_i + k}\right)^{k} \tag{2.9}$$

We use $NB(\mu_i, k)$ to denote the negative binomial distribution with parameters μ_i and k. We also let $dnbinom(y_i; \mu_i, k)$ and $pnbinom(y_i; \mu_i, k)$ denote its PMF and CDF respectively. It can be shown that assuming k fixed and considering the natural parameter $\theta_i = \log\left(\frac{\mu_i}{\mu_i + k}\right)$, $b(\theta_i) = -\log(1 - \exp(\theta_i))$, $a(\phi) = 1/k$, negative binomial is a member of an exponential dispersion family appropriate for discrete variables (a slightly different definition than (2.1); see [1, 20]). By equations (2.4) and (2.5),

$$E(y_i) = \mu_i \tag{2.10}$$

$$V(y_i) = \mu_i + \frac{\mu_i^2}{k} \tag{2.11}$$

2.1.3 Gamma

Suppose y_i has a Gamma distribution with parameters k and μ_i, then its probability density function is

$$f(y_i; \mu_i, k) = \frac{(k/\mu_i)^k}{\Gamma(k)} y_i^{k-1} e^{-\frac{k\, y_i}{\mu_i}} \tag{2.12}$$

A Gamma distribution with parameters k and μ_i is denoted by $Gamma(\mu_i, k)$. Its PDF and CDF are denoted by $dgamma(y_i; \mu_i, k)$ and $pgamma(y_i; \mu_i, k)$, respectively. Parameter k is called the *shape* parameter. Now considering the natural parameter $\theta_i = -1/\mu_i$, $b(\theta_i) = -\log(-\theta_i) = \log(\mu_i)$, and $a(\phi) = 1/k$, Gamma has exponential dispersion form of (2.1). Now, by equations (2.4) and (2.5),

$$E(y_i) = b'(\theta_i) = \frac{-1}{\theta_i} = \mu_i \tag{2.13}$$

$$V(y_i) = b''(\theta_i)a(\phi) = \frac{1}{\theta^2} \times \frac{1}{k} = \frac{\mu_i^2}{k} \tag{2.14}$$

2.2 Zero-Inflated models

In practice, very often, we have excessive zeros in count data, which might not be captured by usual Poisson or negative binomial models. Such data are usually referred to as *zero-*

inflated data. One of the models that has been utilized commonly to describe zero-inflated data is *zero-inflated Poisson (ZIP)*. The zero-inflated Poisson model with parameters λ_i and p, denoted by $ZIP(\lambda_i, p_i)$, is defined by [1, 22]

$$y_i = \begin{cases} 0 & \text{with probability } p_i \\ \text{Poisson}(\lambda_i) & \text{with probability } 1 - p_i \end{cases} \tag{2.15}$$

where λ_i is the mean of the Poisson model and p_i is the probability of excessive zero for ith observation. We denote its PMF and CDF by $dzip(y_i; \lambda_i, p_i)$ and $pzip(y_i; \lambda_i, p_i)$ respectively. Then, unconditional probability distribution is

$$dzip(y_i = 0) = p_i + (1 - p_i)e^{-\lambda_i} \tag{2.16}$$

$$dzip(y_i = j) = (1 - p_i)\frac{e^{-\lambda_i}\lambda_i^j}{j!} \tag{2.17}$$

The mean and variance of a ZIP random variable can be calculated by

$$E(y_i) = \mu_i = (1 - p_i)\lambda_i \tag{2.18}$$

$$V(y_i) = (1 - p_i)\lambda_i \left[1 + p_i\lambda_i\right] \tag{2.19}$$

As it can be seen from the above formulas, $V(y_i) > \mu_i$, so ZIP is another form of overdispersed Poisson.

Note that ZIP does not belong to the exponential dispersion family. It is also not necessary that the explanatory variable describing the λ_i be the same as those describing p_i. The parameters can be modeled by

$$\text{logit}(p_i) = \boldsymbol{Z}\boldsymbol{\gamma} \tag{2.20}$$

$$\log(\lambda_i) = \boldsymbol{X}\boldsymbol{\beta} \tag{2.21}$$

where $\boldsymbol{Z}$ and $\boldsymbol{X}$ are model matrices containing the value of all explanatory variables for p_i and λ_i, and $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are corresponding parameter vectors, respectively. Zero-inflated negative binomial (ZINB) can be defined analogously [1, 8].

2.3 Residuals

2.3.1 Deviance Residuals

Let $l(\boldsymbol{y};\boldsymbol{\mu})$ be the log-likelihood function. A *saturated model* [1, 29] is one in which there are as many estimated parameters as data points. By definition, this will lead to a perfect fit and has the highest log-likelihood among all models. For example, one can easily show that for Poisson, negative binomial, and Gamma regressions, $l(\boldsymbol{y},\boldsymbol{y})$ is the highest achievable log-likelihood and so it is the likelihood for the corresponding saturated model.

Scaled deviance is defined as twice the difference between log-likelihood for saturated model and fitted model. Symbolically, suppose $l\left(\boldsymbol{y};\hat{\boldsymbol{\mu}}\right)$ and $l\left(\boldsymbol{y};\tilde{\boldsymbol{\mu}}\right)$ are the log-likelihood for fitted and saturated model, respectively, then the likelihood ratio statistic is

$$2\left\{l\left(\boldsymbol{y};\tilde{\boldsymbol{\mu}}\right)-l\left(\boldsymbol{y};\hat{\boldsymbol{\mu}}\right)\right\} \tag{2.22}$$

For exponential dispersion family, this has the form of

$$2\sum_{i=1}^{n}\left\{\frac{y_i\tilde{\theta}_i-b(\tilde{\theta}_i)}{a(\phi)}-\frac{y_i\hat{\theta}_i-b(\hat{\theta}_i)}{a(\phi)}\right\} \tag{2.23}$$

where $\hat{}$ and $\tilde{}$ denote the parameters in the fitted and saturated model, respectively. In GLM, usually $a(\phi)=\frac{\phi}{\omega_i}$ for a known weight ω_i, so the likelihood ratio statistic is

$$D^*\left(\boldsymbol{y};\hat{\boldsymbol{\mu}}\right)=\frac{2\sum_{i=1}^{n}\omega_i\left[y_i(\tilde{\theta}_i-\hat{\theta}_i)-b(\tilde{\theta}_i)+b(\hat{\theta}_i)\right]}{\phi}=\frac{D\left(\boldsymbol{y};\hat{\boldsymbol{\mu}}\right)}{\phi} \tag{2.24}$$

The statistics $D^*\left(\boldsymbol{y};\hat{\boldsymbol{\mu}}\right)$ and $D\left(\boldsymbol{y};\hat{\boldsymbol{\mu}}\right)$ are called *scaled deviance* and *deviance*, respectively, and are used as goodness-of-fit (GOF) test. *Deviance residual* is defined as signed square root of the component of $D\left(\boldsymbol{y};\hat{\boldsymbol{\mu}}\right)$, i.e.

$$d_i=\text{sign}(y_i-\hat{\mu}_i)\sqrt{2\left\{\omega_i\left[y_i(\tilde{\theta}_i-\hat{\theta}_i)-b(\tilde{\theta}_i)+b(\hat{\theta}_i)\right]\right\}} \tag{2.25}$$

As it can be seen, $D\left(\boldsymbol{y};\hat{\boldsymbol{\mu}}\right)=\sum_i d_i^2$.

For Poisson, assuming the number of observation, n, is fixed, if the expected counts is large enough, then asymptotically $d_i\sim N(0,1)$ and $D\left(\boldsymbol{y};\hat{\boldsymbol{\mu}}\right)=\sum_i d_i^2\sim\chi^2_{n-p}$, where p is the

number of model parameters in the fitted model [1].

However, it is sometimes challenging to define deviance residuals, particularly, when the model is complex and it is not easy to find the saturated model. For example, for ZIP, it can be shown that Poisson(y_i) is the saturated model for $ZIP(\lambda_i, p)$. So, the deviance residual for ZIP is defined as (see for example [23]) signed square root of the likelihood ratio between the fitted model (zero-inflated Poisson) and the saturated model (Poisson). Table 2.1 summarizes the deviance residuals for different models.

Table 2.1: Deviance residuals for different models

Model	Deviance Residuals
Poisson	$d_i = \text{sign}(y_i - \hat{\mu}_i)\left(2\left\{y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)\right\}\right)^{1/2}$
Negative Binomial	$d_i = \text{sign}(y_i - \hat{\mu}_i)\left(2\left\{y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i + k)\log \frac{y_i+k}{\hat{\mu}_i+k}\right\}\right)^{1/2}$
Gamma	$d_i = \text{sign}(y_i - \hat{\mu}_i)\left(2\left\{-\log \frac{y_i}{\hat{\mu}_i} + \frac{y_i-\hat{\mu}_i}{\hat{\mu}_i}\right\}\right)^{1/2}$
ZIP	$d_i = \text{sign}(y_i - \hat{\mu}_i)\Big(2\Big\{-y_i + y_i \log y_i - \log y_i! - I(y_i = 0)\log\Big[\hat{p}_i + (1-\hat{p}_i)e^{-\hat{\lambda}_i}\Big] - I(y_i > 0)\log\Big[(1-\hat{p}_i) - \hat{\lambda}_i + y_i \log \hat{\lambda}_i - \log y_i!\Big]\Big\}\Big)^{1/2}$

2.3.2 Pearson Residuals

In the GLM context, the *Pearson residual* is the most commonly used measure of goodness of fit, which is defined as

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{V(y_i)}}} \tag{2.26}$$

where $\hat{\mu}_i$ is the fitted value of y_i and $\widehat{V(y_i)}$ is the estimation of variance of y_i. In other words, Pearson residuals are raw residuals scaled by estimation of standard deviation of the response variable. Specifying Pearson residuals for different models are straightforward, with some of the common ones presented in Table 2.2.

Suppose true parameters μ_i and $V(y_i)$ are known, the Pearson residual has mean and standard deviation of 0 and 1. When $\phi = 1$, $X^2 = \sum_i r_i^2$ is the score statistic for comparing fitted model with the corresponding saturated model [29, 43]. In fact, for Poisson model, if μ_i is large enough and the model holds, then asymptotically $r_i \sim N(0,1)$ and $X^2 = \sum_i r_i^2 \sim \chi^2_{n-p}$. X^2 is the well-known *Pearson chi-squared statistic* and is used as a GOF test [29]. As shown in Table 2.2, for Gamma regression, the Pearson residual is related to the shape parameter k (reciprocate for dispersion parameter), so when k is unknown, the scaled version of Pearson residuals $\frac{r_i}{\sqrt{k}} = \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}$ may be used instead.

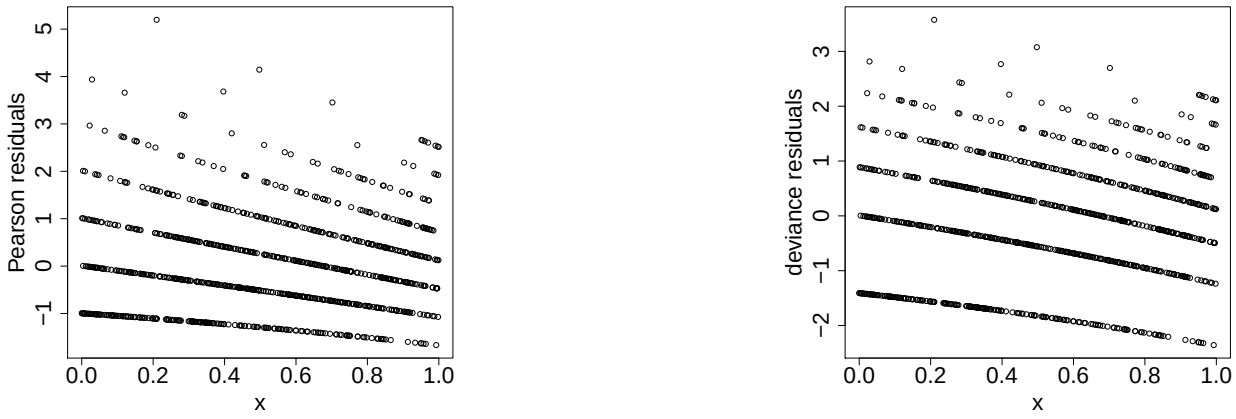
Table 2.2: Pearson residuals for different models

Model	Pearson Residuals
Poisson	$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$
Negative Binomial	$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i + \hat{\mu}_i^2/k}}$
Gamma	$r_i = \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i/\sqrt{k}}$
ZIP	$r_i = \frac{y_i - (1-\hat{p}_i)\hat{\lambda}_i}{\sqrt{(1-\hat{p}_i)\hat{\lambda}_i\left[1+\hat{p}_i\hat{\lambda}_i\right]}}$

2.3.3 Problems with Traditional Residuals

For the normal linear model, the Pearson and deviance residuals are equal and they are exactly normal under the true model. However, Pearson and deviance residuals usually have variance less than 1 since the true mean μ_i is typically unknown, the fitted mean $\hat{\mu}_i$ used instead to compare with y_i, so their distribution is often skewed and non-normally distributed [1, 29, 32]. Theoretically, deviance residual should be more normal than Pearson residual, and if $\phi/\mu_i \to 0$ both Pearson and deviance converge to normal; however, when ϕ/μ_i is large enough, neither of them follow a normal distribution, and the mean and standard deviation for deviance residuals is not necessarily 0 and 1 even the true values μ_i are chosen [15, 32].

In regression models for modeling discrete outcomes, the residuals are far from normality, with residuals aligning nearly parallel curves according to distinct response values, which imposes great challenges for visual inspection. As such, the residual plots for modeling



(a) **(b)**

Figure 2.1: Pearson and deviance residuals for a count data (Poisson regression with $\log(E(\boldsymbol{y})) = \boldsymbol{x}$, where $\boldsymbol{x}$ is a covariate which is uniformly distributed from 0 to 1.

discrete outcome variables give very limited meaningful information for model diagnosis, which renders it of no practical use. For example, in Poisson regression with small mean, the Pearson or deviance residual plot form nearly parallel curves corresponding distinct response values, as demonstrated in Figure 2.1, where the data are generated from a Poisson regression model with a continuous covariate $\boldsymbol{x}$ which is uniformly distributed from 0 to 1. As it can be seen, both Pearson residual and deviance residual plots do not provide much meaningful information for model diagnosis for the fitted model.

2.4 Randomized Quantile Residuals

Randomized quantile residuals [15] for a continuous random variable are defined by taking the probit transformation of the cumulative distribution function (CDF) of the response variable. In discrete case, only some randomization will be added to make the CDF continuous. This can be described symbolically as follows. Let $F(Y;\mu,\phi)$ be the CDF of the random variable Y. In discrete case, let $p(Y;\mu,\phi)$ be the PMF of Y. Consider

$$F^*(Y;\mu,\phi,U) = \begin{cases} F(Y;\mu,\phi) & F \text{ is continuous} \\ F(Y^-;\mu,\phi) + U\, p(Y;\mu,\phi) & F \text{ is disccrete} \end{cases} \tag{2.27}$$

where $F(Y^-;\mu,\phi)$ is the lower limit of F in Y and U is a uniform random variable on $(0,1]$. Then, the randomized quantile residuals [15] are defined as

$$q_i = q(y_i;\hat{\mu}_i,\hat{\phi}_i,u_i) = \Phi^{-1}(F^*(y_i;\hat{\mu}_i,\hat{\phi}_i,u_i)) \tag{2.28}$$

where $\Phi()$ is the cumulative distribution function of standard normal, $\hat{\mu}_i$ is the fitted value for y_i, and u_i is a uniform random variable on $(0,1]$.

If F is continuous, then

$$q_i = \Phi^{-1}\{F(y_i;\hat{\mu}_i,\hat{\phi}_i)\} \tag{2.29}$$

If F is not continuous, let $a_i = \lim_{y\to y_i^-} F(y;\hat{\mu}_i,\hat{\phi}_i)$ and $b_i = F(y_i;\hat{\mu}_i,\hat{\phi}_i)$, then the randomized quantile residual is

$$q_i = \Phi^{-1}(F_i^*), \tag{2.30}$$

where F_i^* is a uniform random variable on the interval $(a_i,b_i]$.

This definition is a special case of "crude residuals" defined by Cox and Snell [11, 15]. As it can be seen from the definition, the randomized quantile residual has a straightforward definition for all distributions. For example, for Poisson, negative binomial, Gamma, and ZIP, see Table 2.3. The only information that is necessary for computing randomized quantile residual is knowing the cumulative distribution function of the response variable, which is a great advantage over deviance residuals, which requires derivation of the saturated model. Nevertheless, in discrete case, the randomized quantile residual depends on the choice of the U and different values for the U lead to different residuals for the same observation. So, researchers may suspect that finding the pattern in the randomized quantile residuals depends heavily on the choice of U. Dunn and Smyth [15] suggested computing and plotting the randomized quantile residuals four times. Then, any pattern in the residuals which is not consistent across the realizations should be ignored. However, when the sample size is relatively large, there might be no need to run them for four times. The reason is because for a given Y_i of the distribution Y, as the sample size increases, there would be more observation with value y_i according to $p(Y_i)$. So, the randomized quantile residual needs to choose more uniform values in any interval $(F(y_i^-),F(y_i)]$, which reduces the chance of having any pattern due to the choice of U. In this dissertation, we will only present one

realization of the randomized quantile residuals in different scenarios. Next, we will show that given true values of the parameters, $q_i \sim N(0,1)$.

Table 2.3: Randomized quantile residuals for different models

Model	Randomized Quantile Residuals
Poisson	$q_i = \Phi^{-1}\Big(ppois(y_i - 1; \hat{\mu}_i) + u_i \cdot dpois(y_i; \hat{\mu}_i)\Big)$
Negative Binomial	$q_i = \Phi^{-1}\Big(pnbinom(y_i - 1; \hat{\mu}_i, \hat{k}) + u_i \cdot dnbinom(y_i; \hat{\mu}_i, \hat{k})\Big)$
Gamma	$q_i = \Phi^{-1}\Big(pgamma(y_i; \hat{\mu}_i, \hat{k})\Big)$
ZIP	$q_i = \Phi^{-1}\Big(pzip(y_i - 1; \hat{\mu}_i, \hat{p}_i) + u_i \cdot dzip(y_i; \hat{\mu}_i, \hat{p}_i)\Big)$

Theorem 2.4.1. *Let $F(Y)$ denotes the true CDF of Y and in discrete case, assume also that $p(Y)$ denotes PMF of Y. If $U \sim Unif\,(0,1]$, then*

$$q = q(Y; \mu, \phi, U) \sim N(0,1)$$

Proof. First, we show that $F^*(Y,U)$ in equation (2.27) is a uniform random variable on $(0,1]$ which is equivalent of proving that $F^*(Y,U)$ has the same CDF as the uniform random variable on $(0,1]$. So, it is enough to show that for any $0 < t \leq 1$, $P(F^*(Y,U) \leq t) = t$. If F is continuous, then because F is non-decreasing

$$P(F^*(Y,U) \leq t) = P(F(Y) \leq t) = P(Y \leq F^{-1}(t)) = F(F^{-1}(t)) = t,$$

where, $F^{-1}(t) = \inf\{X : F(X) \geq t\}$. Now, assume that Y is a discrete random variable with values $y_1, y_2, \ldots$. Then, for $0 < t \leq 1$, let k be the largest i such that $F(y_i) \leq t$, then

$$P(F^*(Y,U) \leq t) = \sum_{i=1}^{k} P(F(y_i^-) < F^*(Y,U) \leq F(y_i)) + P(F(y_k) < F^*(Y,U) \leq t) \quad (2.31)$$

By (2.30), because $F^*(Y,U)$ is a uniform random variable on each $\big(F(y_i^-), F(y_i)\big]$, then $F(y_i^-) < F^*(Y,U) \leq F(y_i)$ if and only if $Y = y_i$. So,

$$P(F(y_i^-) < F^*(Y,U) \leq F(y_i)) = P(Y = y_i)$$

For evaluating $P(F(y_k) < F^*(Y,U) \leq t)$, note that because k is the maximum i such that $F(y_i) \leq t$, $F(y_k) < F^*(Y,U) \leq t$ implies that $Y = y_{k+1}$. Because $F^*(Y,U) \leq t$, so $U \leq \frac{t-F(Y^-)}{p(Y)}$, thus

$$
\begin{aligned}
&P(F(y_k) < F^*(Y,U) \leq t) \\
&\qquad = P(F(y_k) < F^*(Y,U) \leq t) \\
&\qquad = P(Y = y_{k+1} \,\&\, U \leq \frac{t - F(Y^-)}{p(Y)}) \\
&\qquad = P(Y = y_{k+1}) \cdot P(U \leq \frac{t - F(Y^-)}{p(Y)} | Y = y_{k+1}) \\
&\qquad = P(Y = y_{k+1}) \cdot P(U \leq \frac{t - F(y_{k+1}^-)}{p(y_{k+1})}) \\
&\qquad = P(Y = y_{k+1}) \cdot P(U \leq \frac{t - F(y_k)}{p(y_{k+1})}) \quad (\text{because} F(y_{k+1}^-) = F(y_k)) \\
&\qquad = p(y_{k+1}) \cdot \frac{t - F(y_k)}{p(y_{k+1})} \\
&\qquad = t - F(y_k)
\end{aligned}
$$

So, by equation (2.31)

$$
\begin{aligned}
&P(F^*(Y,U) \leq t) \\
&\qquad = \sum_{i=1}^{k} P(F(y_i^-) < F^*(Y,U) \leq F(y_i)) + P(F(y_k) < F^*(Y,U) \leq t) \\
&\qquad = \sum_{i=1}^{k} p(y_i) + t - F(y_k) \\
&\qquad = F(y_k) + t - F(y_k) = t
\end{aligned}
$$

So, as it can be seen, F^* has the same CDF as uniform, resulting that it is indeed uniformly distributed. Hence,

$$P(\Phi^{-1}(F^*(Y,U)) \leq t) = P(F^*(Y,U) \leq \Phi(t)) = \Phi(t)$$

So, $\Phi^{-1}(F^*(Y,U)$ has the same CDF as the standard normal distribution, indicating that $q = q(Y; \mu_i, \phi_i, U) \sim N(0,1)$. □

2.4.1 Illustrative Example

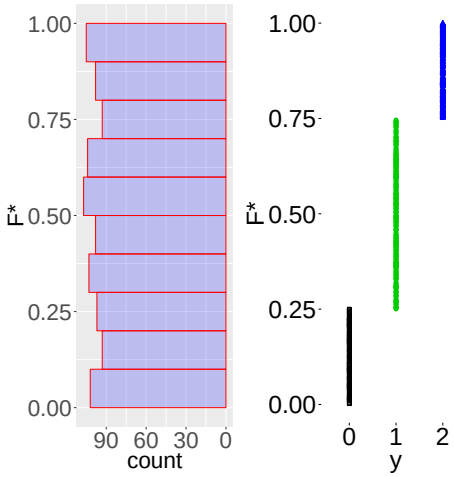
As an example, suppose that Y has a binomial distribution with $n = 2$ and $p = .5$. We expect that almost a quarter of observations be 0, half of the observations 1, and a quarter of observations 2. So, $F^*(Y, U)$ assigns uniform numbers in (0, .25] to almost a quarter of data (when 0 is observed), uniform numbers in (.25, .75] to half of data (when 1 is observed), and uniform numbers in (.75, 1] for the last quarter of data (when 2 is observed). Thus, $F^*(Y, U)$ should be uniformly distributed on (0 ,1]. To illustrate, we simulate 1000 data points from this distribution and compute $F^*(y_i)$. The result is depicted in Figure 2.2a, indicating $F^*(y_i)$ is indeed uniformly distributed. On the other hand, suppose that we wrongly fit the following model to the data:

Y	0	1	2
$p(Y)$	.1	.8	.1

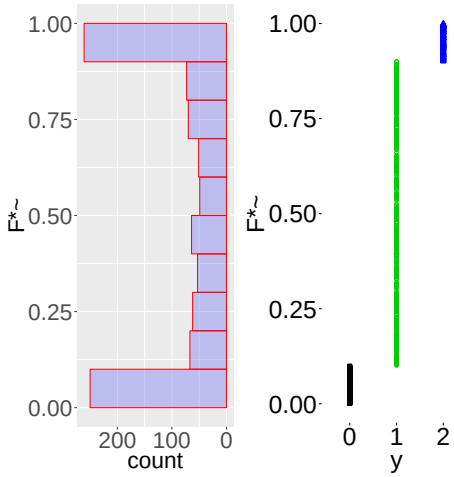
(2.32)

Now, if we compute $\tilde{F}^*(y_i)$ for this model, all observations that are 0 (around a quarter of data) will be assigned uniformly to the interval (0, .1]. All observations that are 1 (around half of data) will be uniformly assigned to the interval (.1, .9], and finally all observations that are 2 (around a quarter of data) will be assigned uniformly to the interval (.9, 1]. So, $\tilde{F}^*$ will have heavy tails in comparison to the middle of the data, indicating that $\tilde{F}^*$ is not uniformly distributed and so the model is wrong (Figure 2.2b).

Furthermore, if we compute randomized quantile residuals for the models above (both true and wrong model), we will see that randomized quantile residuals for the true model is indeed normally distributed, if the binomial model with $n = 2$ and $p = .5$ is chosen (Figure 2.3), but the residuals are not normal for the wrong model based on the model defined by PMF (2.32) (Figure 2.4).

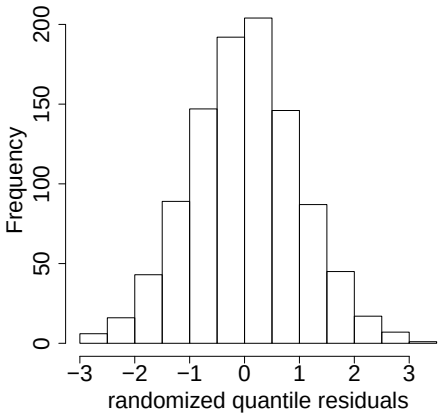


(a) F^* for true model (binomial with $n = 2$ and $p = .5$)



(b) $\tilde{F}^*$ for wrong model (model fitted based on 2.32)

Figure 2.2: F^* for true model in the left and $\tilde{F}^*$ for the wrong model in the right



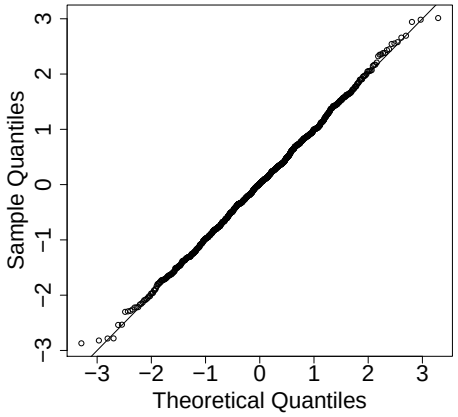
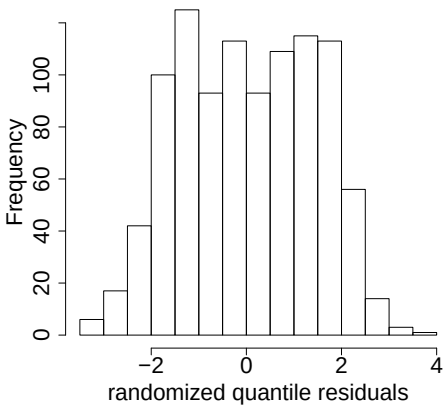


Figure 2.3: Histogram and QQ-plot for randomized quantile residuals when the fitted model is the true model



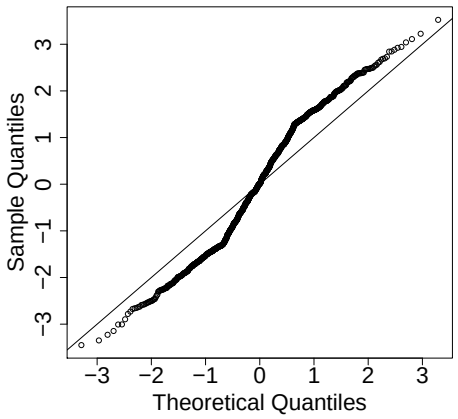


Figure 2.4: Histogram and QQ-plot for randomized quantile residuals when the fitted model is the wrong model

2.5 Normality Tests for Randomized Quantile Residuals

As we have seen in previous Sections, under the true model, one expects the residuals to be roughly normal and approximately independent distributed with a mean of zero and some constant variance. In this Section, we review some of the most important normality tests that can test whether the data is normal or not. For all the information and definition in this Section, we follow [46].

2.5.1 Wilk-Shapiro Test

The Wilk-Shapiro test utilizes the null hypothesis principle to check whether a sample $x_1, x_2, \cdots, x_n$ comes from a normally distributed population. For the set of observation x_i, let $x_{(i)}$ be the ith order statistic and $w_i = E(x_{(i)})$ be the expected value of them based on the assumption that $x_i \sim N(0,1)$. Let also $\boldsymbol{V}$ be the covariance matrix of $x_{(i)}$. Suppose $\boldsymbol{x}^T = \left(x_{(1)}, \ldots, x_{(n)}\right)^T$ and $= \left(w_1, \ldots, w_n^T\right)$. Let

$$b = \mathbf{a}^T \boldsymbol{x}, \tag{2.33}$$

where $\mathbf{a}^T = \dfrac{\mathbf{w}^T \mathbf{V}^{-1}}{\sqrt{\mathbf{w}^T \mathbf{V}^{-1} \mathbf{V}^{-1} \boldsymbol{w}}}$ and $\mathbf{a}^T \boldsymbol{a} = \mathbf{1}$. Then the *Wilk-Shapiro* [41] test statistic is defined as

$$W = \frac{b^2}{(n-1)s^2}, \tag{2.34}$$

where s^2 is the sample variance.

The b above is (up to a constant) generalized least square regression of $\boldsymbol{x}$ on $\boldsymbol{w}$, which is the best linear unbiased estimate of σ, variance of the population. One of the issues with this test is to calculate the elements of $\boldsymbol{V}$ and hence $\boldsymbol{a}$. These values are known exactly only for samples up to size 20, and are estimated for greater sample sizes.

2.5.2 Shapiro-Francia Test

As an approximation to Wilk-Shapiro test, the *Shapiro-Francia* [40] test statistic is defined as

$$W = \frac{(\mathbf{a}^* \boldsymbol{x})^2}{(n-1)\, s^2}, \tag{2.35}$$

where

$$\mathbf{a}^* = \frac{\mathbf{w}^T}{\sqrt{\mathbf{w}^T \boldsymbol{w}}} \tag{2.36}$$

As it can be seen W' is the squared Pearson correlation between $\mathbf{a}^*$ and x. A substantial benefit of Shapiro-Francia comparison to Wilk-Shapiro is that it only needs the expected values of the order statistics be known, and it is not necessary to calculate $\boldsymbol{V}$; however, the necessity for calculating expected values also makes this test cumbersome.

2.5.3 EDF Tests

Empirical distribution function (EDF) tests are those GOF tests which compares the empirical and hypothetical distribution functions.

Empirical distribution function (EDF) of a sample, $F_n(x)$, is defined as

$$F_n(x) = \begin{cases} 0 & x < x_{(1)} \\ i/n & x_{(i)} \leq x < x_{(i+1)} \qquad i = 1, \ldots, n-1 \\ 1 & x_{(n)} \leq x \end{cases} \tag{2.37}$$

EDF tests reject normality when the discrepancies between the EDF and the hypothetical cumulative distribution function of normal distribution

$$p_{(i)} = \Phi(\frac{x_{(i)} - \hat{\mu}}{\hat{\sigma}}) \tag{2.38}$$

are too large.

The Kolmogorov-Smirnov Test

Let

$$D^+ = \max_{i=1,\ldots,n} \left\{ i/n - p_{(i)} \right\} \tag{2.39}$$

$$D^- = \min_{i=1,\ldots,n} \left\{ p_{(i)} - (i-1)/n \right\} \tag{2.40}$$

Then, the Kolmogorov-Smirnov test statistic is defined as the maximum discrepancies between the EDF and the $p_{(i)}$, i.e.

$$D = \max \left\{ D^+, D^- \right\} \tag{2.41}$$

Lilliefors Test

Lilliefors [26] was the first who raises the question of using EDF tests for composite hypotheses, and he gave a table of critical values for D based on simulation. The p-value for this test is computed by the Dallal-Wilkinson formula [12], which seems to be reliable when the p-value is smaller than 0.1. If the computed p-value turns out to be greater than 0.1, then the modification

$$D^* = \left(\sqrt{n} - 0.01 + 0.85/\sqrt{n} \right) D \tag{2.42}$$

can be used to compute the p-value [44].

Anderson-Darling Test

Anderson and Darling [3] proposed a class of EDF GOF tests by

$$n \int_{-\infty}^{+\infty} \left[F_n(x) - F(x) \right]^2 \psi(F(x)) d(F(x)) \tag{2.43}$$

where $F(x)$ is the hypothesized distribution function and $\psi(F(x))$ is a weighting function.

Later on, they used $\psi(p) = [p(1-p)]^{-1}$ as the weighting function in (2.43) to define *Anderson-Darling* test statistic [4] as

$$A^2 = -n - n^{-1} \sum_{i=1}^{n} [2i-1] \left[\log \left(p_{(i)} \right) + \log \left(1 - p_{(n-i+1)} \right) \right] \tag{2.44}$$

A set of critical values for all sample sizes can be obtained by Stephens modification [45]

$$A^{2*} = \left(1.0 + 0.75/n + 2.25/n^2 \right) A^2 \tag{2.45}$$

Cramer-von Mises Test

Using $\psi(F(x)) = 1$ in equation (2.43) results in *Cramer-von Mises* test statistic

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(p_{(i)} - \frac{2i-1}{2n} \right)^2 \tag{2.46}$$

with the modification

$$W^{2*} = (1.0 + 0.5/n)W^2 \tag{2.47}$$

for computing the p-value [45]. [1]

Among all these tests, Wilk-Shapiro test has the highest power [37, 42]. So, in next chapter, for brevity, we only show the results for Wilk-Shapiro test, however, other methods will give almost the same results. For conducting different normality tests, we used R-package "nortest" [17].

One problem with these normality tests is that in replicated experiments, their p-value is very slightly different than uniform distribution, even though the data itself simulated from a normal distribution. Take Wilk-Shapiro test, for example; we can see from figure 2.5 that the last column of the histogram is lower than the others. The same thing happens to other normality tests and the last column of their histogram is a little lower than the others. In fact, in next chapter, when we use wilk-Shapiro test, we will have the same issue every time. Since it is not a serious problem, it can be ignored.

[1] See also [48, 49]

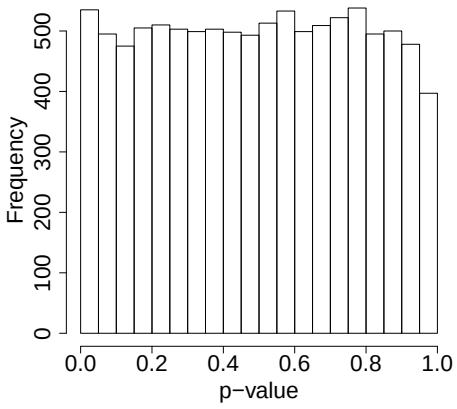


Figure 2.5: P-value from the Wilk-Shapiro test for normal data

Chapter 3

Simulation Studies

In this chapter, we investigate the performance of randomized quantile residuals and compare them with deviance and Pearson residuals using simulated datasets. We consider three different cases: in Section 3.1, non-linearity in the covariate; in Section 3.2, overdispersion; and in Section 3.3, zero-inflation.

3.1 Non-Linearity in the Covariate

We evaluate the performance of linearity tests based on the randomized quantile residuals in comparison with other types of residuals in scenarios where the functional form of a covariate is misspecified. To investigate the ability of detecting the functional form of a covariate, we considered the following models:

- **Model 1:** $\eta_i = \beta_0 + \beta_1 x_i^2$ (3.1a)
- **Model 2:** $\eta_i = \beta_0 + \beta_1 \exp(x_i)$ (3.1b)
- **Model 3:** $\eta_i = \beta_0 + \beta_1 \sin(x_i)$ (3.1c)

where x_i, $i = 1, \cdots, n$ is a continuous covariate from uniform distribution on some interval. Model 1 represent scenario where a quadratic term should be specified and Model 2 represents scenario where a covariate being exponentiated. Model 3 indicates η_i and x_i are related non-linearly following a sin function. The response variable under consideration for these models follow Poisson, negative binomial, and Gamma, which will be presented in the following Scenarios.

Scenario 1: Poisson

First, we focus on distinguishing between linear and quadratic effect in Poisson regression with log-link function. For this purpose, we suppose the sample size $n = 1000$ and we simulate a covariate $\boldsymbol{x}$ from $Uniform(-1.2, 1.2)$. Then, we assume that $\eta_i = x_i^2$ and so $\mu_i = \exp(\eta_i)$. Now, we sample $y_i \sim Poisson(\mu_i)$ and fit two models;

- **True model:**

$$y|x \sim Poisson(\exp(\beta_0 + \beta_1 x^2)) \tag{3.2}$$

- **Wrong model:**

$$y|x \sim Poisson(\exp(\beta_0 + \beta_1 x)) \tag{3.3}$$

Note that here, our main focus is to distinguish between linear and non-linear effect in covariate, so we are not concerned about hierarchical rule, however, the results are the same if we include a linear part in our true model as well. Then, for each model, we calculate different kinds of residuals and we plot them against $\boldsymbol{x}$ and against fitted values. We observe that the plots against the covariate is slightly easier to interpret than that of fitted values. For brevity, we don't include the plots against fitted values, nor we do include residuals versus $\boldsymbol{\eta}$. The plots for different residuals for each model against the covariate are presented in Figure 3.1. As we can see from the results, Pearson and deviance residuals fail to distinguish between these two models. Again, for each of them, we have some parallel lines, each of which corresponds to one of the distinct values of $\boldsymbol{y}$. On the other hand, as it can be seen, from 3.1f the randomized quantile residuals for Poisson model with linear effect suggest that there is a quadratic trend in the residuals, so we should add a quadratic term as well. Moreover, 3.1e also shows that there is no problem with quadratic Poisson model, and that one can fit the model very well. We also tried the same thing with other non-linear functions instead of x^2, such as $sin(x)$, $\exp(x)$, and $\log(x)$, and we could draw the same conclusion. The codes and plots are available upon request.

In normal regression, people often look at the QQ-plot of the residuals to determine if they are normal or not. We want to see if the same thing happens for different residuals. Figure 3.2 depicts the QQ-plot for all three residuals for two models. Figures 3.2a - 3.2d

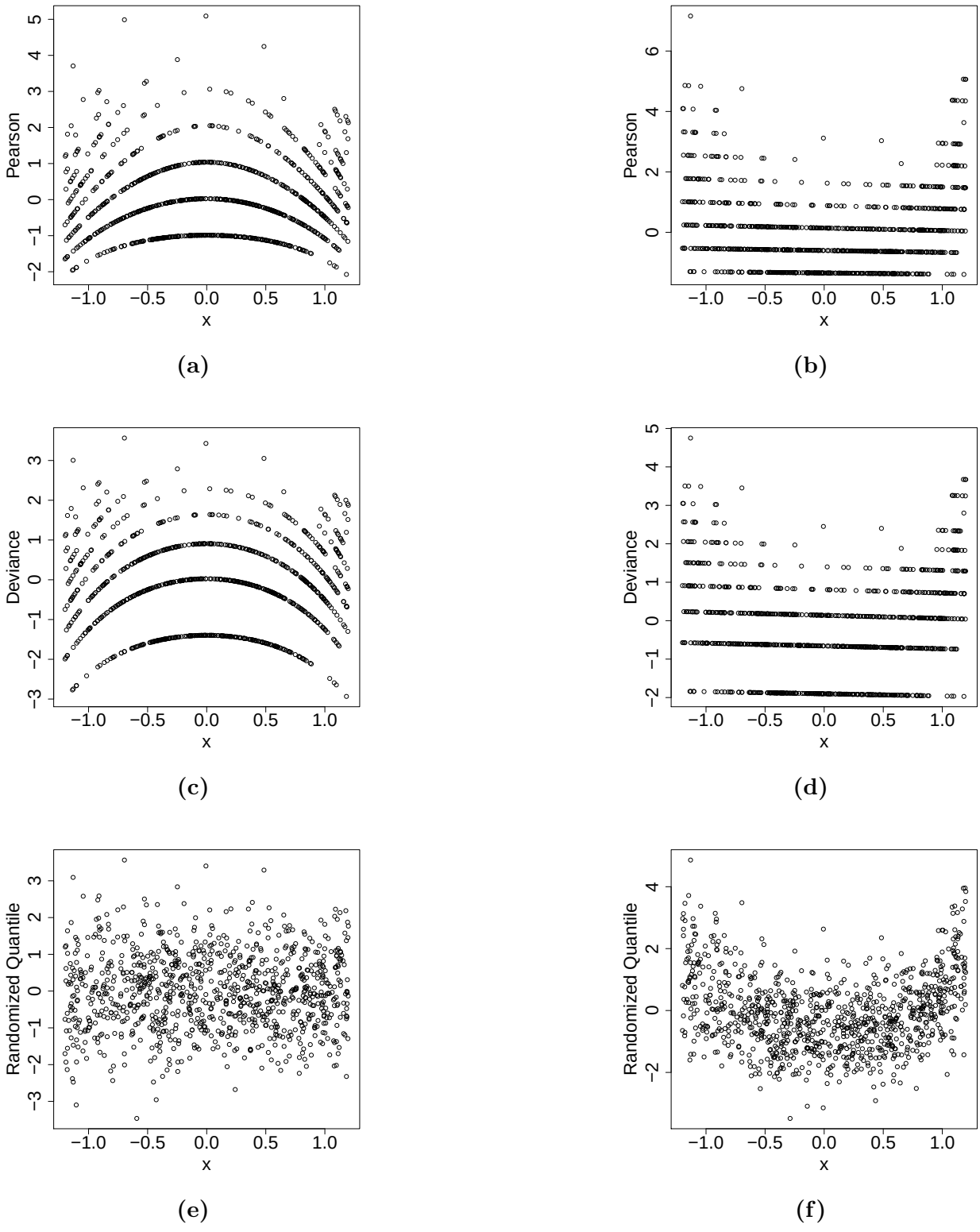


Figure 3.1: Pearson, deviance, and randomized quantile residuals for two models; left panel: $y|x \sim Poisson(\exp(\beta_0 + \beta_1 x^2))$ (true model) and right panel: $y|x \sim Poisson(\exp(\beta_0 + \beta_1 x))$ (wrong model)

show that regardless of the model, neither Pearson, nor deviance follow normal distribution and these residuals fail to choose the true model. On the other hand, Figure 3.2f confirms that the Poisson model with linear effect can not fit the data very well, and Figure 3.2e shows that quadratic Poisson fits the data well. So, randomized quantile residuals can correctly distinguish the true model from the wrong model.

We already demonstrated the superior performance of randomized quantile residual than others for a single dataset. To confirm this finding, we replicate the previous experiment 10000 times. For each of the simulated dataset, we fit Poisson with linear and quadratic effect to each of them. Then, we compute different residuals. To see which of them is normal, we applied different normality tests to the residuals, but we only present here the results from the Wilk-Shapiro test, as it has the highest power among normality tests. The result from other tests are almost the same and will be provided upon request. We know that under the true model, the residual should be normally distributed, so the p-value for the normality should be uniformly distributed. However, shown in Figures 3.3a - 3.3d, even when the fitted model is the true model, the Pearson and deviance residuals are not normal, even though it has been suggested that Pearson is asymptotically normal. On the other hand, results from the randomized quantile residual (Figure 3.3f) shows that if the model is not correct, the randomized quantile residual is not normal. Figure 3.3e confirms that the distribution of the p-value for normality test is uniform, which confirms that the true model has a quadratic form in the covariate effect. So, the randomized quantile residuals perform superior to the Pearson and deviance residuals. The results corresponding to other non-linear functional form of the covariate effect are almost the same and are available upon request.

Scenario 2: Negative Binomial

Now, we want to investigate the difference between linear and non-linear effect in negative binomial model with log-link function. For this purpose, similar to Poisson case, we assume the sample size $n = 1000$ and simulate a covariate $\boldsymbol{x} \sim \mathit{Uniform}(-1.5, 1.5)$. Now, let $\eta_i = x_i^2$ and $\mu_i = \exp(\eta_i)$ (log-link function). Then, we sample $y_i \sim NB(\mu_i, k = 2)$, where k is the reciprocal for the dispersion paremter. We fit two models using "glm.nb" from the R-package "MASS" [47];

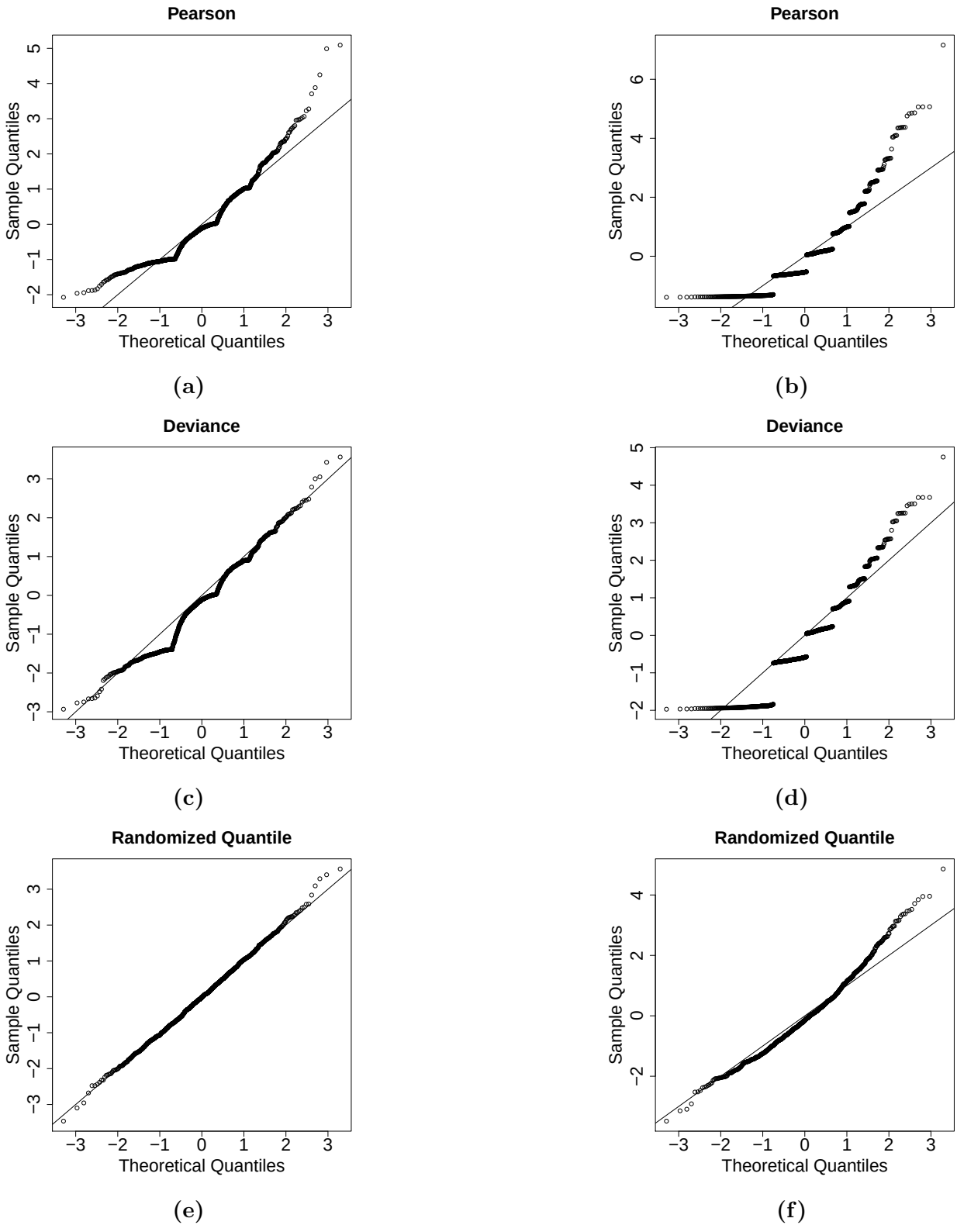


Figure 3.2: QQ-plot for Pearson, deviance, and randomized quantile residuals for two models; left panel: $y|x \sim Poisson(\exp(\beta_0 + \beta_1 x^2))$ (true model) and right panel: $y|x \sim Poisson(\exp(\beta_0 + \beta_1 x))$ (wrong model)

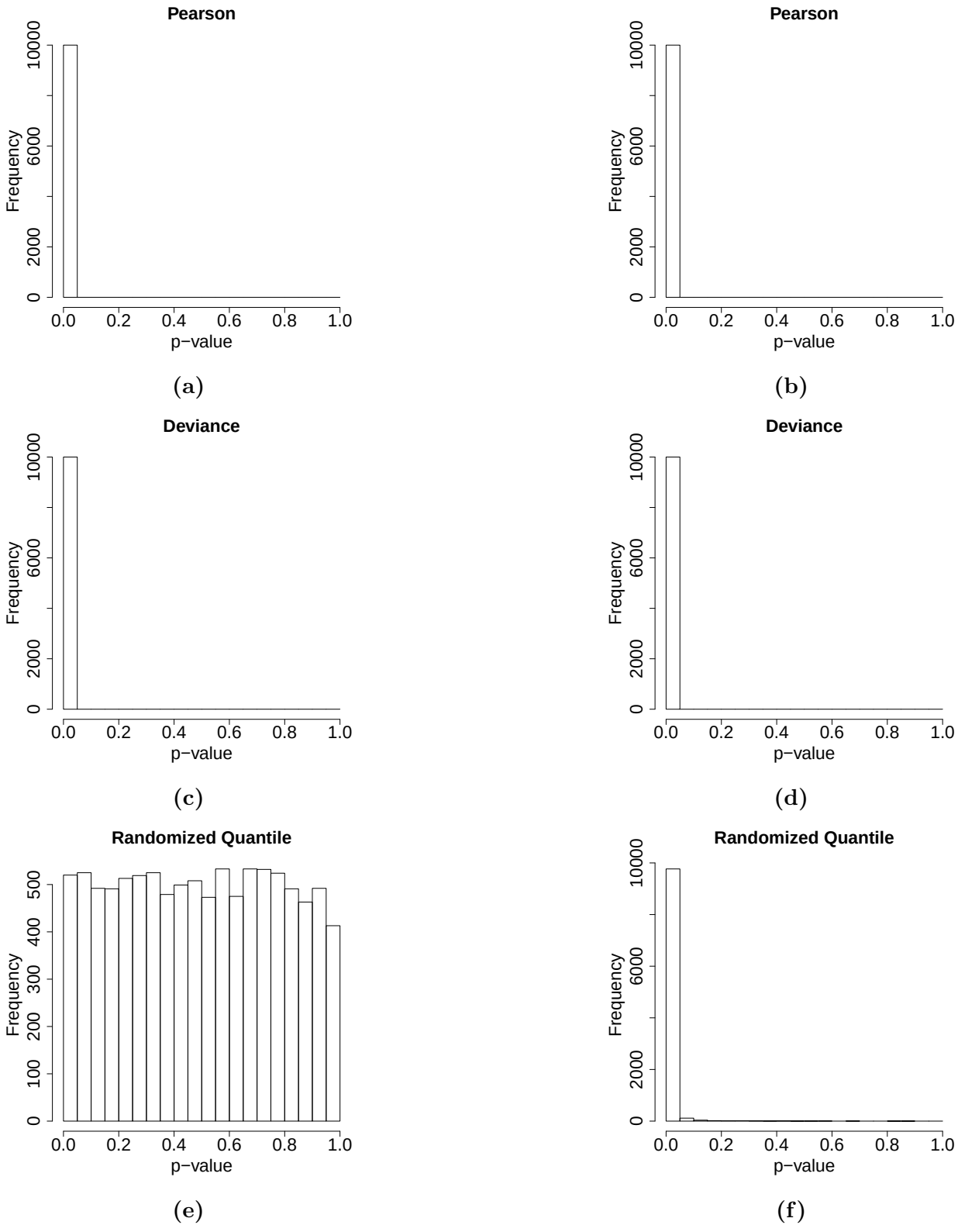


Figure 3.3: The p-value from Wilk-Shapiro test for Pearson, deviance, and randomized quantile residuals for two models; left panel: $y|x \sim Poisson(\exp(\beta_0 + \beta_1 x^2))$ (true model) and right panel: $y|x \sim Poisson(\exp(\beta_0 + \beta_1 x))$ (wrong model)

- **True model:**

$$y|x \sim NB(\exp(\beta_0 + \beta_1 x^2), k) \tag{3.4}$$

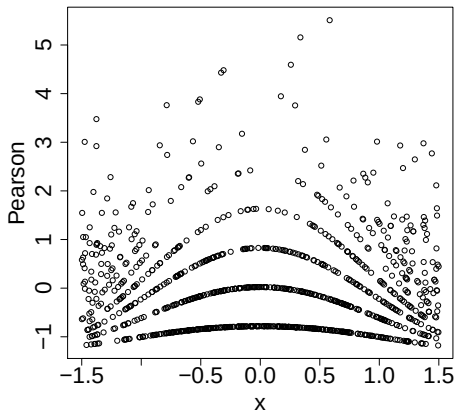
- **Wrong model:**

$$y|x \sim NB(\exp(\beta_0 + \beta_1 x), k) \tag{3.5}$$

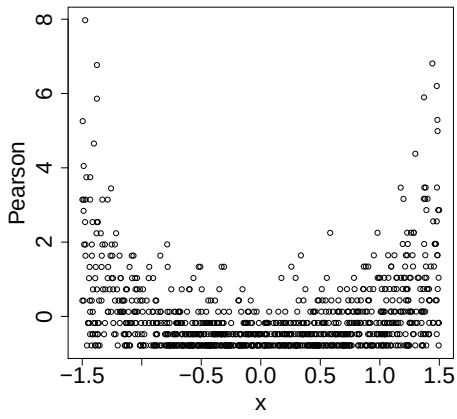
For each model, we compute different kinds of residuals and we plot them against $\boldsymbol{x}$ and against fitted values. Again here, we only discuss those that are plotted against $\boldsymbol{x}$ (Figure 3.4). As one can see from the results, Pearson and deviance residuals could not give much information on which model should be used. Figure 3.4f shows that there is a quadratic pattern in randomized quantile residuals for negative binomial with linear effect and so linear negative binomial does not fit the data well. However, Figure 3.4e substantiates that the quadratic negative binomial model can be considered as the true model as the residuals are scattered randomly.

We also present the QQ-plots of the residuals for easier visual inspection. The QQ-plot for different residuals are depicted in Figure 3.5. As it can be seen from Figures 3.5a - 3.5d, Pearson and deviance fail to distinguish the true model from the wrong model, hence failing to help in model adequacy checking. But, Figure 3.5e and 3.5f shows that randomized quntile residual chooses model with quadratic effect in covariate in comparison to linear one, because there are few data points in tail area that can not be captured by linear effect model as good as quadratic one.

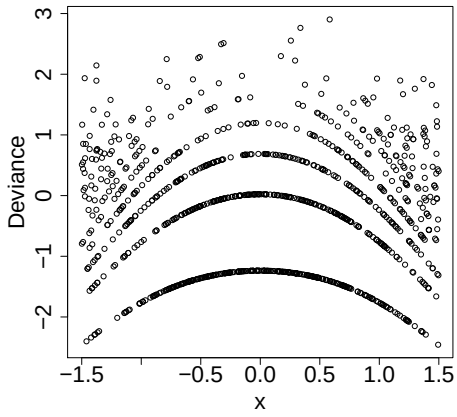
Similar to Poisson case, here, we are going to replicate our experiment 10000 times. To do so, we simulate 10000 datasets, for each of which we fit two models; linear and quadratic effect negative binomial. Then, after computing different residuals for each of them, we apply different normality tests to investigate on the normality of them. Again, we only present the results from the Wilk-Shapiro test (Figures 3.6a - 3.6f). From the histofras, we can see that the Pearson and deviance residuals are not normal irrespective of the model we used. Figures 3.6e and 3.6f show that the randomized quantile residual can detect the true model; as it can be seen, the p-values of Wilk-Shapiro test for randomized quantile residuals are not uniformly distributed and so the model can not be chosen, whereas results from the quadratic model confirms that the p-value is almost uniformly distributed. So, the quadratic model should be



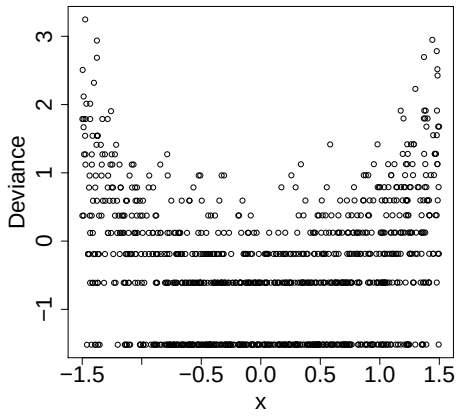
(a)



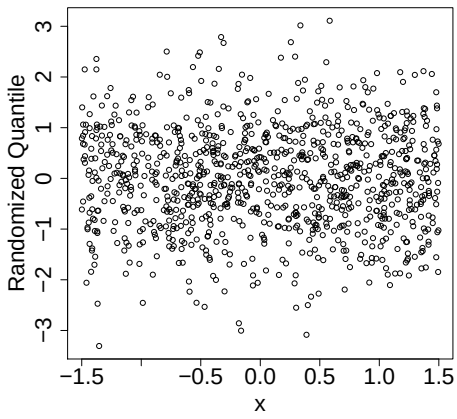
(b)



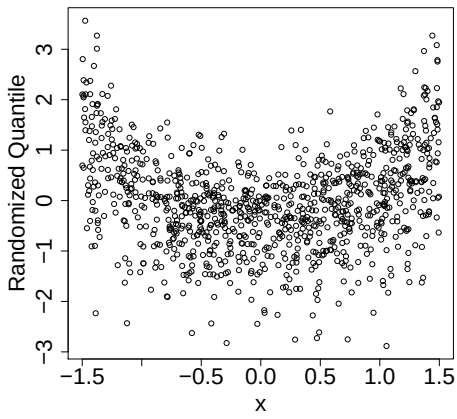
(c)



(d)



(e)



(f)

Figure 3.4: Pearson, deviance, and randomized quantile residuals for two models; left panel: $y|x \sim NB(\exp(\beta_0 + \beta_1 x^2), k)$ (true model) and right panel: $y|x \sim NB(\exp(\beta_0 + \beta_1 x), k)$ (wrong model)

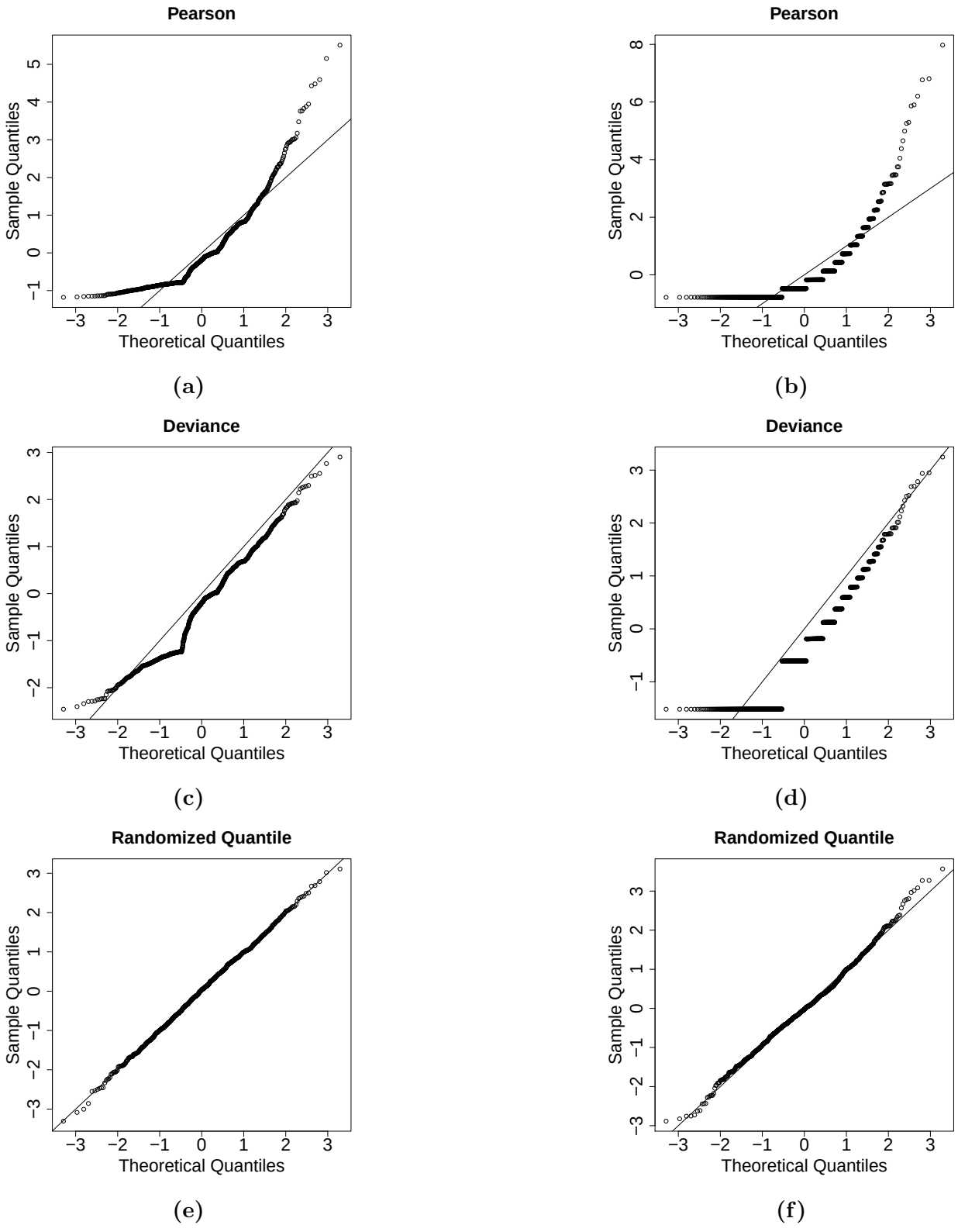


Figure 3.5: QQ-plot for Pearson, deviance, and randomized quantile residuals for two models; left panel: $y|x \sim NB(\exp(\beta_0 + \beta_1 x^2), k)$ (true model) and right panel: $y|x \sim NB(\exp(\beta_0 + \beta_1 x), k)$ (wrong model)

chosen as true model. The results from other normality tests and other non-linear functions are almost the same and will be provided upon request. The only problem of randomized quantile residual is that as shown in the Figure 3.6e, there is a slight increasing trend in p-value for randomized quantile residuals. Although the trend is not very substantial, there are cases where this trend can be problematic, especially when the mean is not large. The reason is because we use the estimation of dispersion parameter to compute the randomized quantile residuals. This is the so-called optimistic bias, which occurs when the actual observations appear to be more predictable by the model.

Scenario 3: Gamma

Gamma regression is one of the most used applicable models in practice. So, aside from count data, we decided to investigate the performance of randmized quantile residuals on Gamma regression. There are three most commonly used link function for Gamma regression: 1- the inverse link $g(\mu_i) = \frac{1}{\mu_i}$, 2- the log link $g(\mu_i) = \log \mu_i$, and 3- the identity link $g(\mu_i) = \mu_i$. Here, we only consider the log link to make it more consistent with other models that we use in this dissertation. Similar to Poisson and negative binomial model, we want to see if the randomized quantile residual can determine linear versus non-linear effect in covariate in Gamma regression. For this purpose, we choose the sample size of $n = 1000$ and we simulate a covariate $\boldsymbol{x}$ from $Uniform(-1.5, 1.5)$. Then, we assume that $\eta_i = x_i^2$ and so $\mu_i = \exp(\eta_i)$. Now, we sample $y_i \sim Gamma(\mu_i, k = 2)$ (with shape parameter 2). Then, we fit two models;

- **True model:**

$$y|x \sim Gamma(\exp(\beta_0 + \beta_1 x^2), k) \tag{3.6}$$

- **Wrong model:**

$$y|x \sim Gamma(\exp(\beta_0 + \beta_1 x), k) \tag{3.7}$$

For estimating the shape parameter, we use the R-package "MASS" [47]. We compute all residuals for both models and plot them against $\boldsymbol{x}$ (Figures 3.7a - 3.7f). As it can be seen from the plots, Figure 3.7b indicates that Pearson residuals can show non-linear trend in linear effect Gamma model, but they cannot distinguish if the quadratic model is the

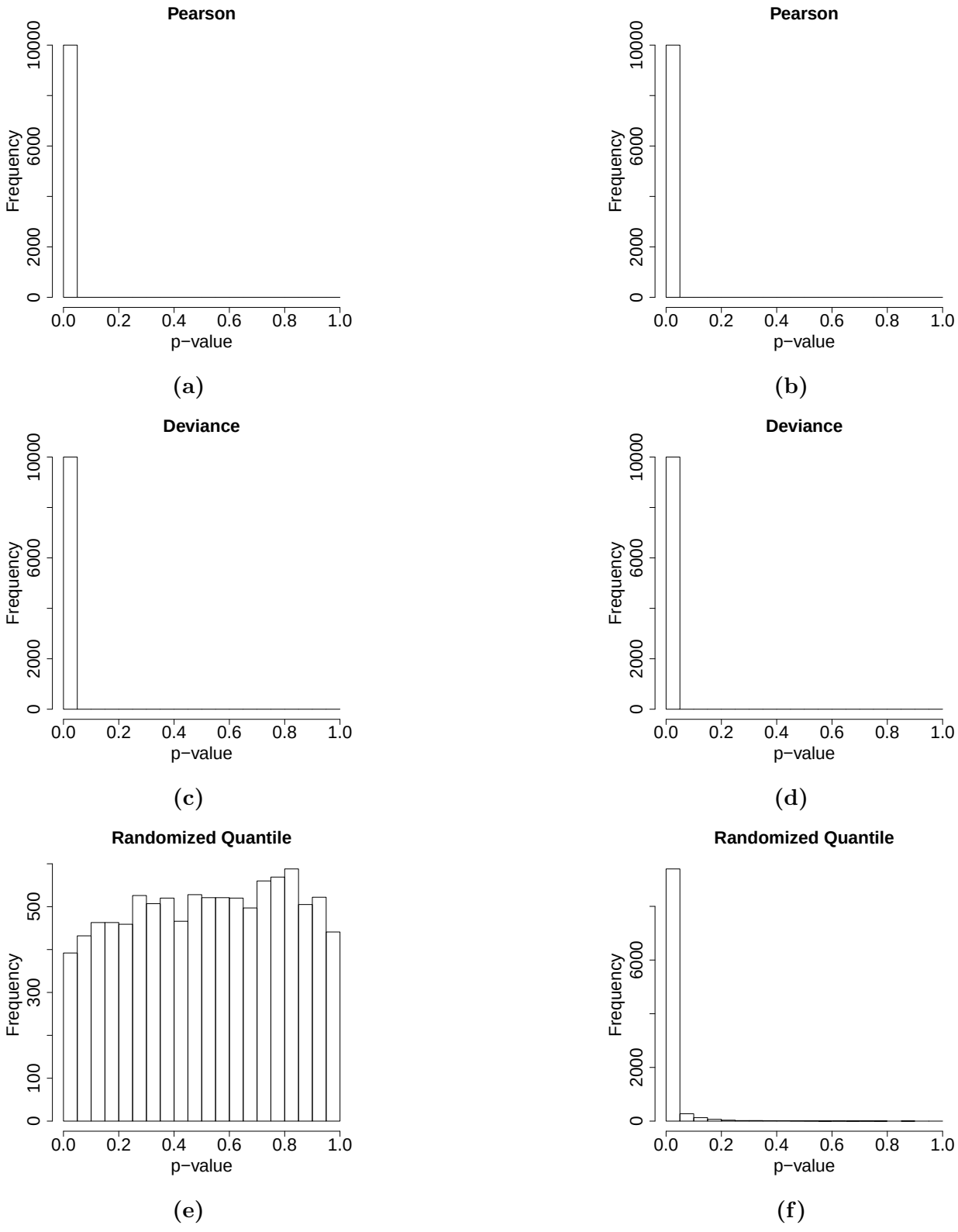
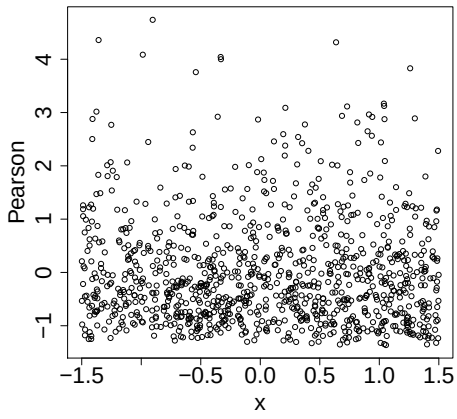


Figure 3.6: The p-value from Wilk-Shapiro test for Pearson, deviance, and randomized quantile residuals for two models; left panel: $y|x \sim NB(\exp(\beta_0 + \beta_1 x^2), k)$ (true model) and right panel: $y|x \sim NB(\exp(\beta_0 + \beta_1 x), k)$ (wrong model)

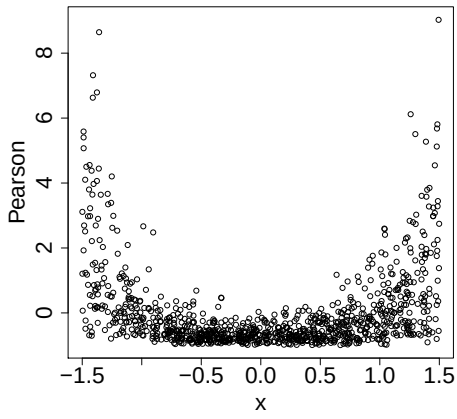
correct one, because as it can be seen from Figure 3.7a, Pearson residuals are not symmetric ranging from -1 to 3, indicating that they are not normal. Figures 3.7c and 3.7d show that in this case, deviance residuals appear to work very well in detecting the true model. They can show that there is a non-linear trend in the residual plots for Gamma model with linear effect, and so the model with linear effect cannot be true. The plots show there is nothing against normality in deviance and residuals for quadratic Gamma model, so they can truly choose the true model. As shown in Figures 3.7e and 3.7f, in this case, randomized quantile residual works well too in distinguishing the linear and non-linear effects in coavriate. There is a quadratic trend in randomized quantile residual for linear effect model, indicating this model does not fit the data, but the randomized quantile residuals for quadratic Gamma model indicate that the normality assumption of residuals is not violated, and so this model can be chosen as true model. The same result can be drawn using other non-linear functions such as $sin(x)$, $\exp(x)$, and $\log(x)$.

QQ-plots for three kinds of residuals are depicted in Figure 3.8. Figures 3.8a and 3.8b show that Pearson residuals are not normal even when the fitted model is true model. Figure 3.8d shows that deviance residual for wrong model is not normal. From Figure 3.8c, we can see although deviance residuals appear to be normal for true model, they are not $N(0, 1)$ as one expects. This makes it hard to visually check if the model is true or not. However, in Figure 3.8f, randomized quantile residuals are not normally distributed and so the model can not be true. But, from Figure 3.8e, randomized quantile residuals for quadratic model are normal, so they can choose the true model comparison with the wrong model. Again in this case, randomized quantile residuals have superior performance than Pearson and deviance.

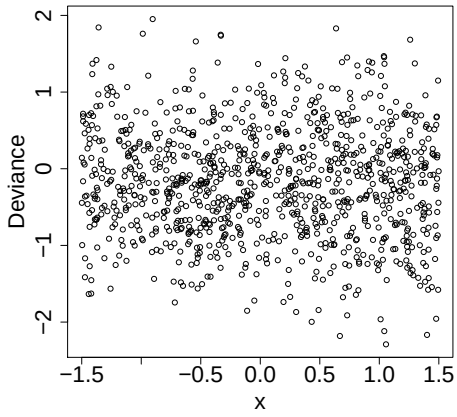
Next, we replicate 10000 datasets from Gamma distribution similar to the previous case, and we fit linear and quadratic effect Gamma models to each of them. Then, we compute different residuals for each of them and we apply normality tests to them to see if they are normal. The histogram of p-values for Wilk-Shapiro test is depicted in Figures 3.9a - 3.9f. Pearson residuals completely fail to distinguish between the two models. The deviance and randomized quantile residuals can correctly choose which model is correct and which model is wrong. When the model is not the true model, neither of the residuals are normal, but when the quadratic model is chosen, then the p-values for both residuals are uniformly distributed,



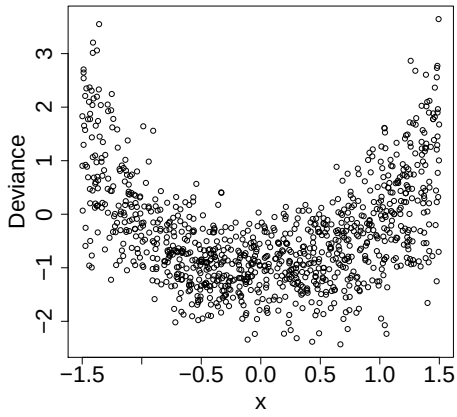
(a)



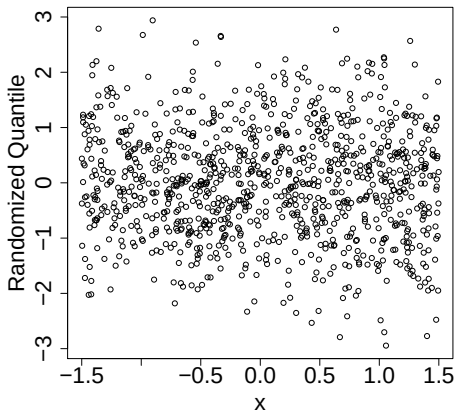
(b)



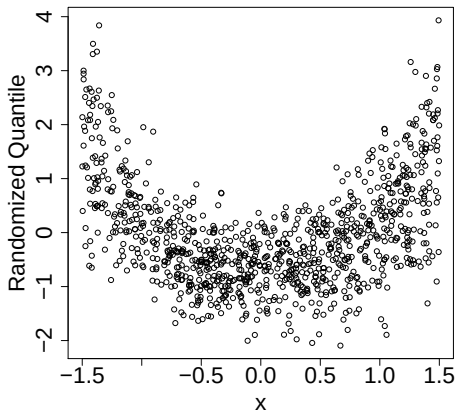
(c)



(d)

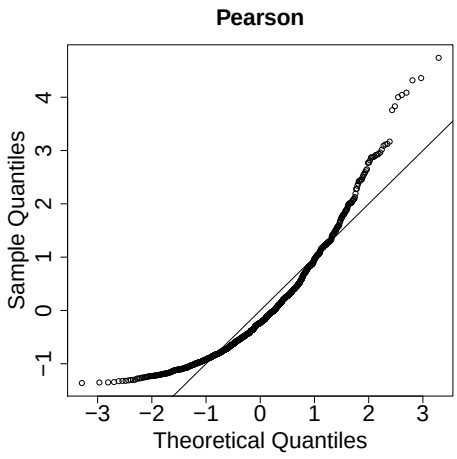


(e)

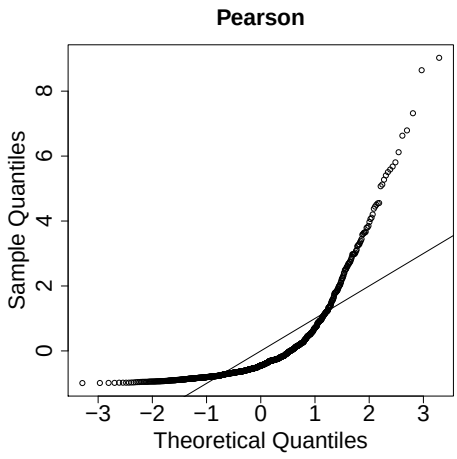


(f) model $y|x \sim Gamma(k, \exp(\beta_0 + \beta_1 x))$

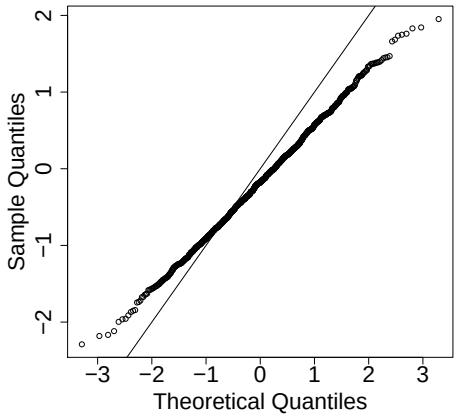
Figure 3.7: Pearson, deviance, and randomized quantile residuals for two models; left panel: $y|x \sim Gamma(k, \exp(\beta_0 + \beta_1 x^2))$ (true model) and right panel: $y|x \sim Gamma(k, \exp(\beta_0 + \beta_1 x))$ (wrong model)



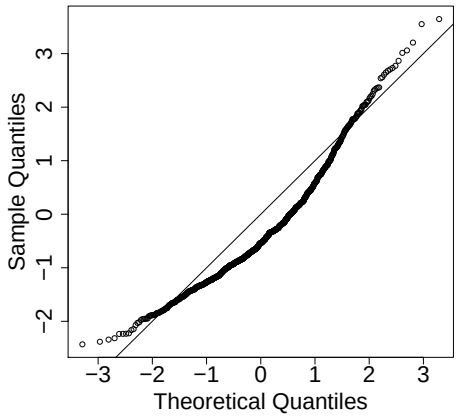
(a)



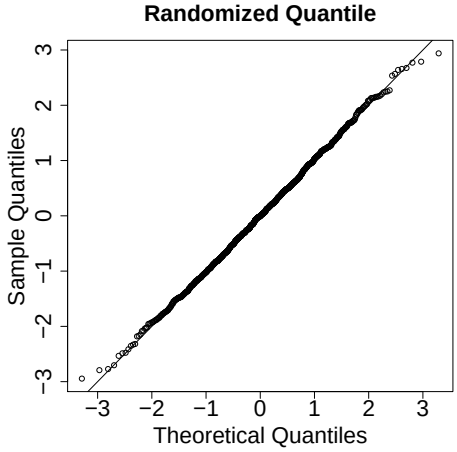
(b)



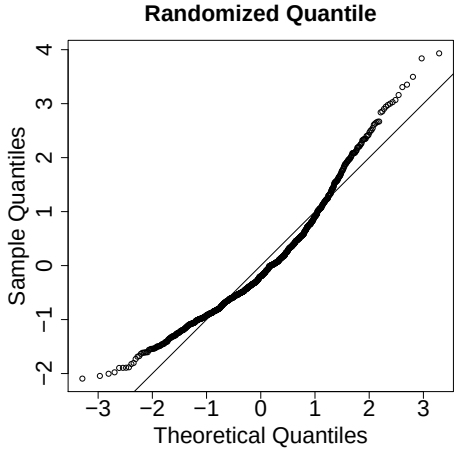
(c)



(d)



(e)



(f) model $y|x \sim Gamma(k, \exp(\beta_0 + \beta_1 x))$

Figure 3.8: Pearson, deviance, and randomized quantile residuals for two models; left panel: $y|x \sim Gamma(k, \exp(\beta_0 + \beta_1 x^2))$ (true model) and right panel: $y|x \sim Gamma(k, \exp(\beta_0 + \beta_1 x))$ (wrong model)

showing that the model is indeed the true model. The results for functions other than x^2 and other normality tests are almost the same and are available upon request.

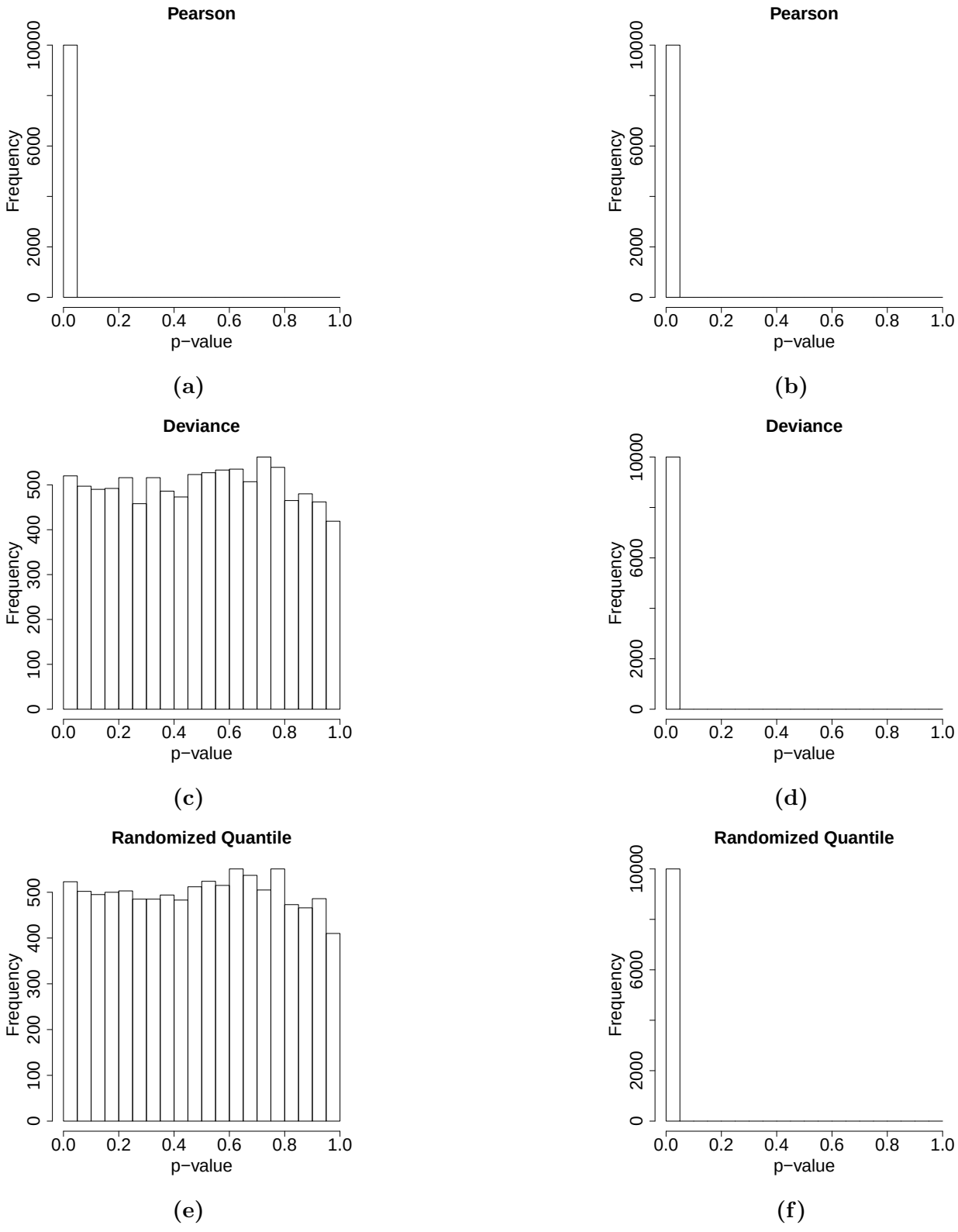


Figure 3.9: P-value from the Wilk-Shapiro test for Pearson, deviance, and randomized quantile residuals for two models; left panel: $y|x \sim Gamma(k, \exp(\beta_0 + \beta_1 x^2))$ (true model) and right panel: $y|x \sim Gamma(k, \exp(\beta_0 + \beta_1 x))$ (wrng model)

3.2 Overdispersion Diagnosis

In this Section, we want to investigate if any of the residuals can tell if the data is overdispersed or not. So, we choose the sample size $n = 1000$ and simulate a covariate $\boldsymbol{x} \sim Uniform(-1.5, 1.5)$. Now, let $\eta_i = \beta_0 + \beta_1 x_i$, where $\beta_0 = 1$ and $\beta_1 = 2$, and let $\mu_i = \exp(\eta_i)$ (log-link function). Then, we simulate $\boldsymbol{y}$ from $NB(\mu_i, k = 2)$ and fit two models;

- **True model:**

$$y|x \sim NB(\exp(\beta_0 + \beta_1 x), k) \tag{3.8}$$

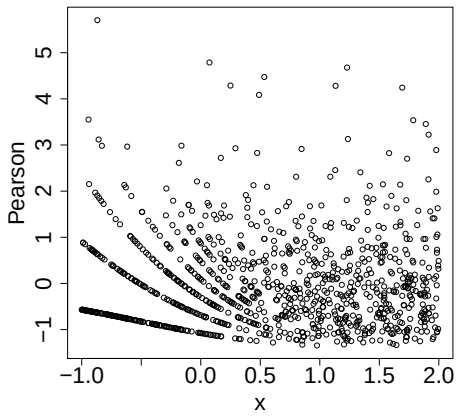
- **Wrong model:**

$$y|x \sim Poisson(\exp(\beta_0 + \beta_1 x), k) \tag{3.9}$$

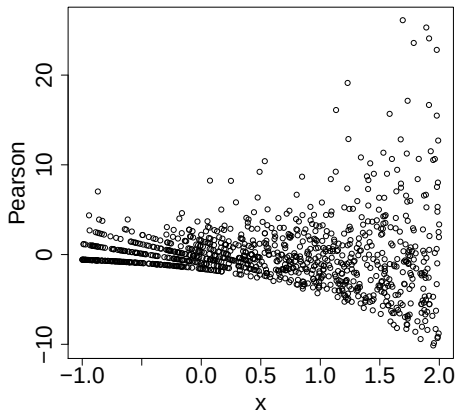
Again, for fitting negative binomial model, we use "glm.nb" from the R-package "MASS". After fitting these models, we calculate different kinds of residuals. Figures 3.10 depicts plots of residuals versus $\boldsymbol{x}$. Although Pearson and deviance residuals suggest that the negative binomial might fit the dataset better, they cannot tell if the negative binomial model can fit well the data or not. By Figures 3.10e and 3.10f, we can see that the randomized quantile residual can suggest us correctly that the data is overdispersed and Poisson model can not fit the data very well, whereas there is no problem in randomized quantile residual for negative binomial and that model indeed fits the data very well.

QQ-plots of the above residuals are depicted in Figure 3.11. Pearson residuals fail to detect the true model and overdispersion. Deviance residuals for true model has better QQ-plot than wrong model, yet still there is a little problem with their QQ-plot for true model. The QQ-plots for randomized quantile residuals show that randomized quantile residual for true model is normal nd for wrong model is far from normality, suggesting that randomized quantile residuals can detect overdispersion better than deviance and Pearson residuals.

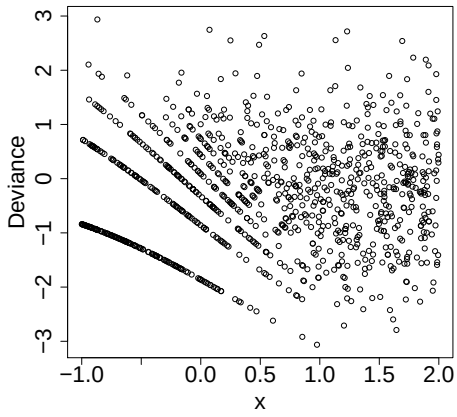
In order to replicate our experiment 10000 times, we simulate 10000 datasets from negative binomial similar to the single dataset experiment. For each of these data sets, we fit two models; Poisson and negative binomial, and then, for each of the two models, we compute different residuals and the p-value from the Normality tests. The histogram for Wilk-Shapiro p-value is depicted in Figures 3.6a - 3.6f. Pearson and deviance residuals are not normal



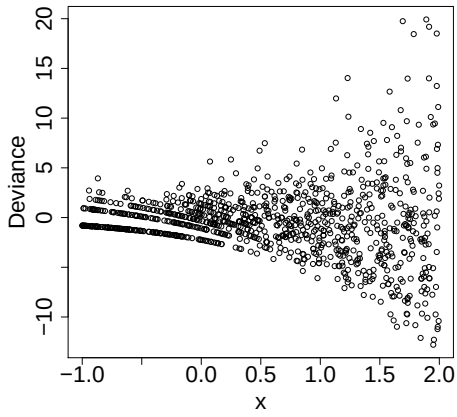
(a)



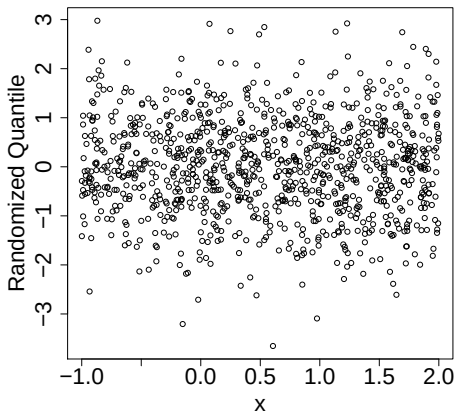
(b)



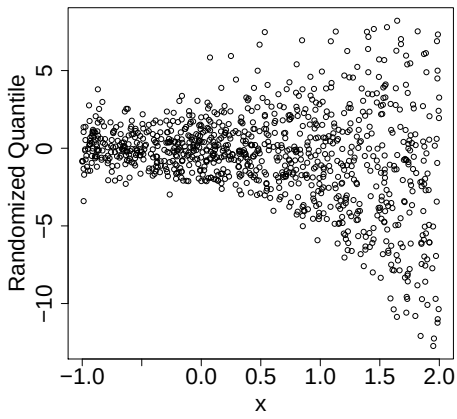
(c)



(d)



(e)



(f)

Figure 3.10: Pearson, deviance, and randomized quantile residuals for two models; left panel: $y|x \sim NB(\exp(\beta_0 + \beta_1 x), k)$ (true model) and right panel: $y|x \sim Poisson(\exp(\beta_0 + \beta_1 x))$ (wrong model)

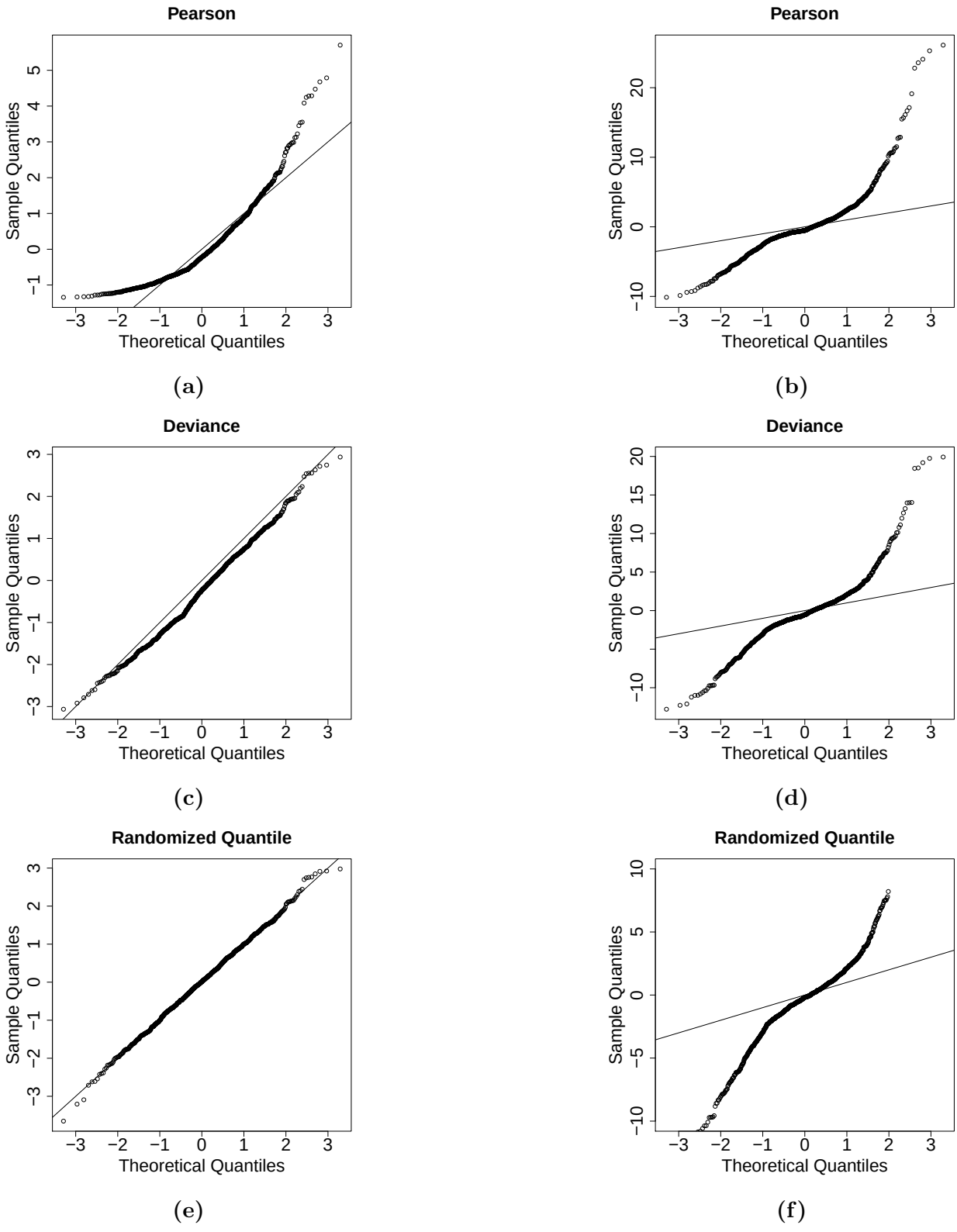


Figure 3.11: QQ-plot for Pearson, deviance, and randomized quantile residuals for two models; left panel: $y|x \sim NB(\exp(\beta_0 + \beta_1 x), k)$ (true model) and right panel: $y|x \sim Poisson(\exp(\beta_0 + \beta_1 x))$ (wrong model)

irrespective of the model we used. Figure 3.6e shows that the p-values for randomized quantile residuals are not uniformly distributed and so the model can not be chosen, whereas results from the randomized quantile residual (Figure 3.6f) confirms that the p-value is almost uniformly distributed if the true model is chosen. The results from other normality tests are almost the same and will be provided upon request.

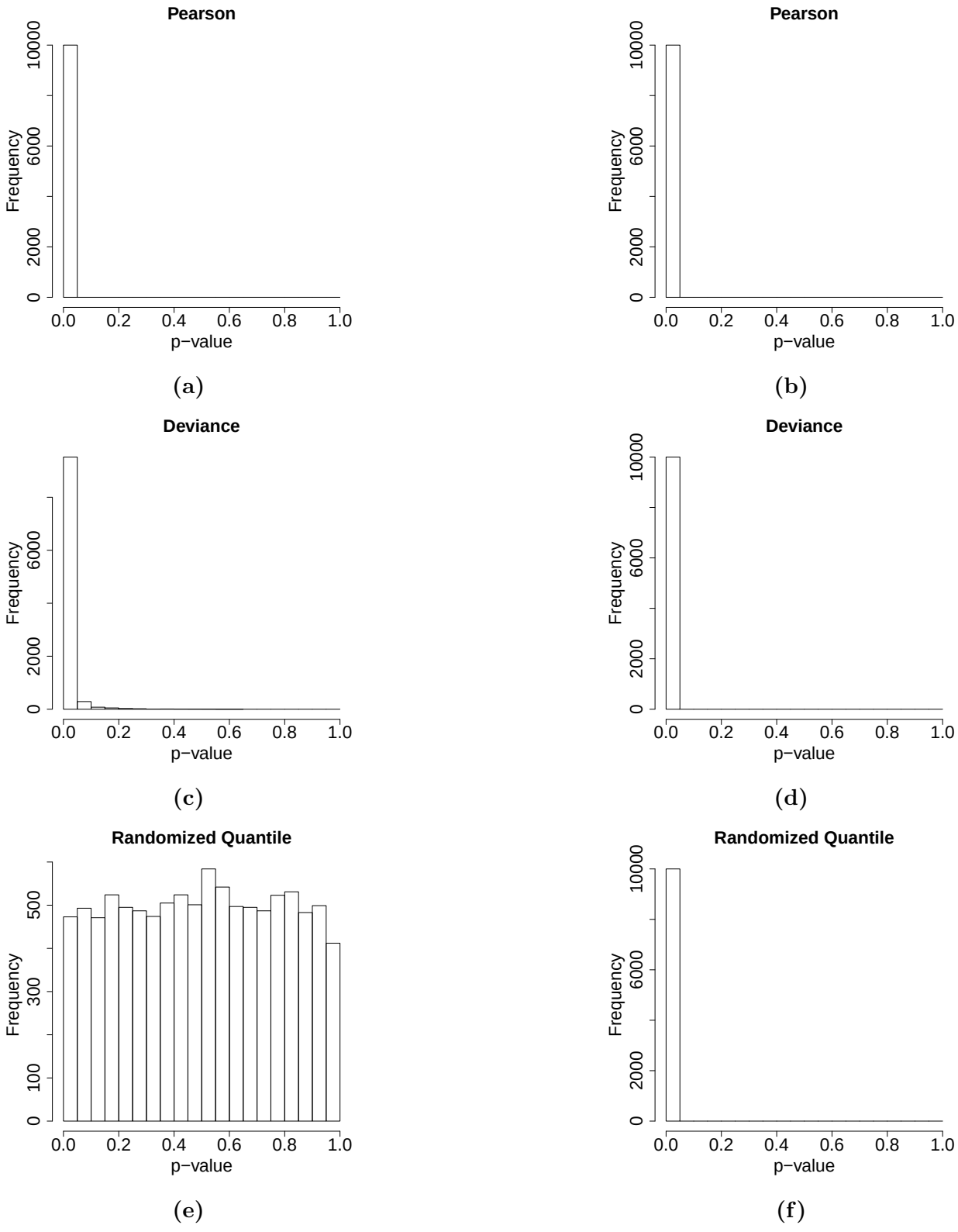


Figure 3.12: P-value from the Wilk-Shapiro test for Pearson, deviance, and randomized quantile residuals for two models; left panel: $y|x \sim NB(\exp(\beta_0 + \beta_1 x), k)$ (true model) and right panel: $y|x \sim Poisson(\exp(\beta_0 + \beta_1 x))$ (wrong model)

3.3 Zero-Inflation Diagnosis

For simulation, we simulate a covariate $\boldsymbol{x}$ of size 1000 from uniform distribution on $(-1, 2)$. Then, we assume $\lambda_i = \exp\{1+2x_i\}$ and simulate y_i from $ZIP(\lambda_i, p = .3)$. We fit two models;

- **True model:**

$$y|x \sim ZIP(\exp(\beta_0 + \beta_1 x), p) \tag{3.10}$$

- **Wrong model:**

$$y|x \sim Poisson(\exp(\beta_0 + \beta_1 x)) \tag{3.11}$$

For fitting ZIP model, we use the function "zeroinfl" from R-package "pscl" [51]. Different residuals for these two models are depicted in Figure 3.13. "pscl" provides only Pearson residuals and for other residuals, we write codes by ourselves. As it can be seen from Figures 3.13a - 3.13d, neither Pearson nor deviance residual can help distinguish between the true model and the wrong model because they do not appear normal and have different variance for different x. Figures 3.13f shows that the Poisson can not fit the data very well and the curve line in that diagram shows that there might be an inflation in one observation of data. Figure 3.13f shows that there is no problem with randomized quantile residuals and so randomized quantile residual can choose the true model.

The QQ-plots of different residuals are also depicted in Figure 3.14. Both Pearson and deviance completely fail to identify the true model from the wrong model. The QQ-plot for randomized quantile residuals works perfectly well in detecting zero-inflation. The QQ-plot for the randomized quantile residuals for true model confirms their normality, while the QQ-plot for the wrong model is not normal. Thus, randomized quantile residual works superior to the deviance and Pearson in detecting zero-inflation in count data.

For replication, we generate 10000 datasets from ZIP similar to the single dataset experiment. For each of these datasets, we fit two models; true model: $y|x \sim ZIP(\exp(\beta_0+\beta_1 x), p)$ and wrong model: $y|x \sim Poisson(\exp(\beta_0 + \beta_1 x))$. After calculating different residuals for them, we apply Wilk-Shapiro test. The histogram for Wilk-Shapiro p-value is depicted in Figures 3.15a - 3.15f. As it can be seen the Pearson and deviance residuals are not normal even when the true model is fitted. From Figure 3.15f, we can conclude that since the p-values

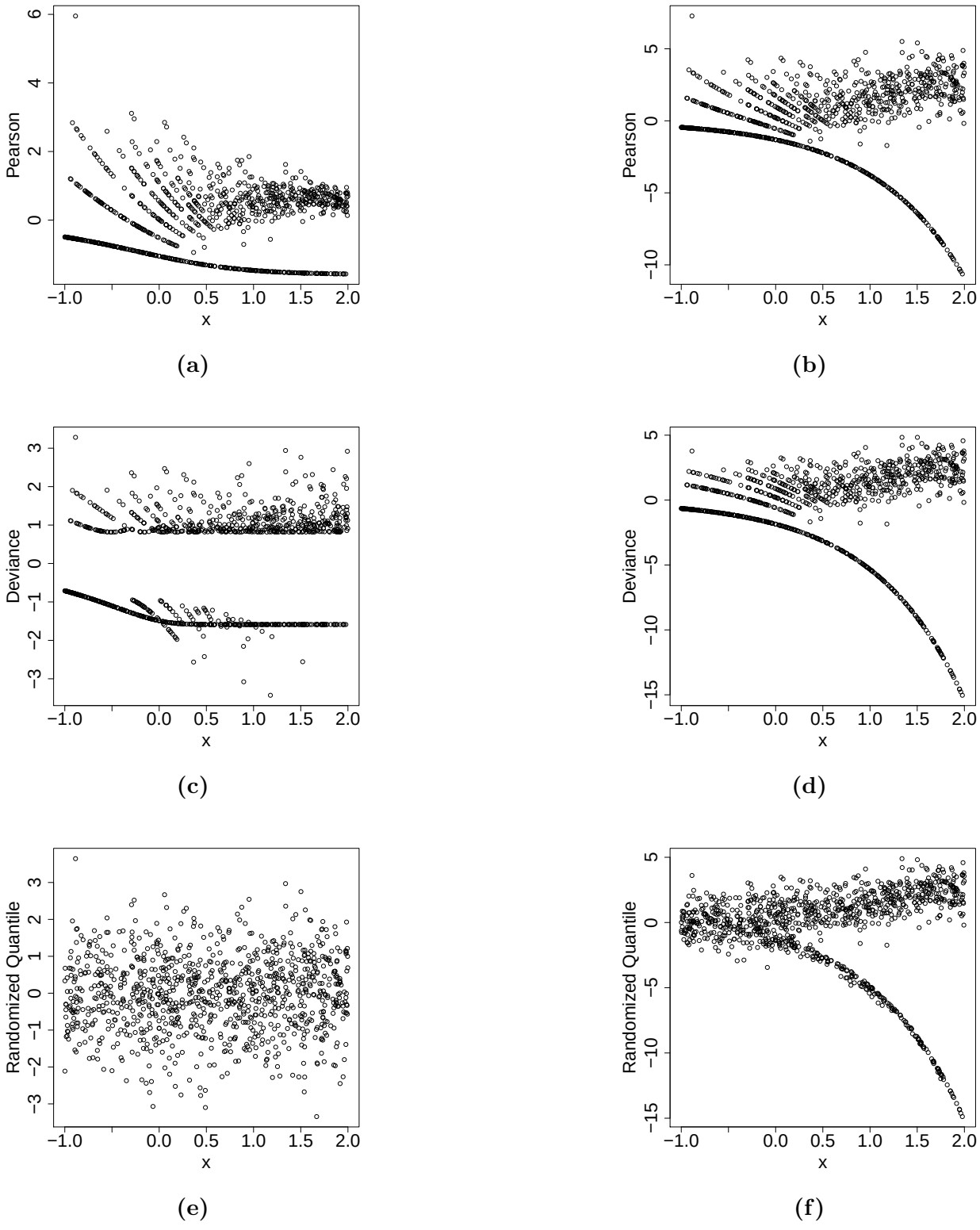


Figure 3.13: Pearson, deviance, and randomized quantile residuals for two models; left panel: $y|x \sim ZIP(\exp(\beta_0+\beta_1 x))$ (true model) and right panel: $y|x \sim Poisson(\exp(\beta_0+\beta_1 x))$ (wrong model)

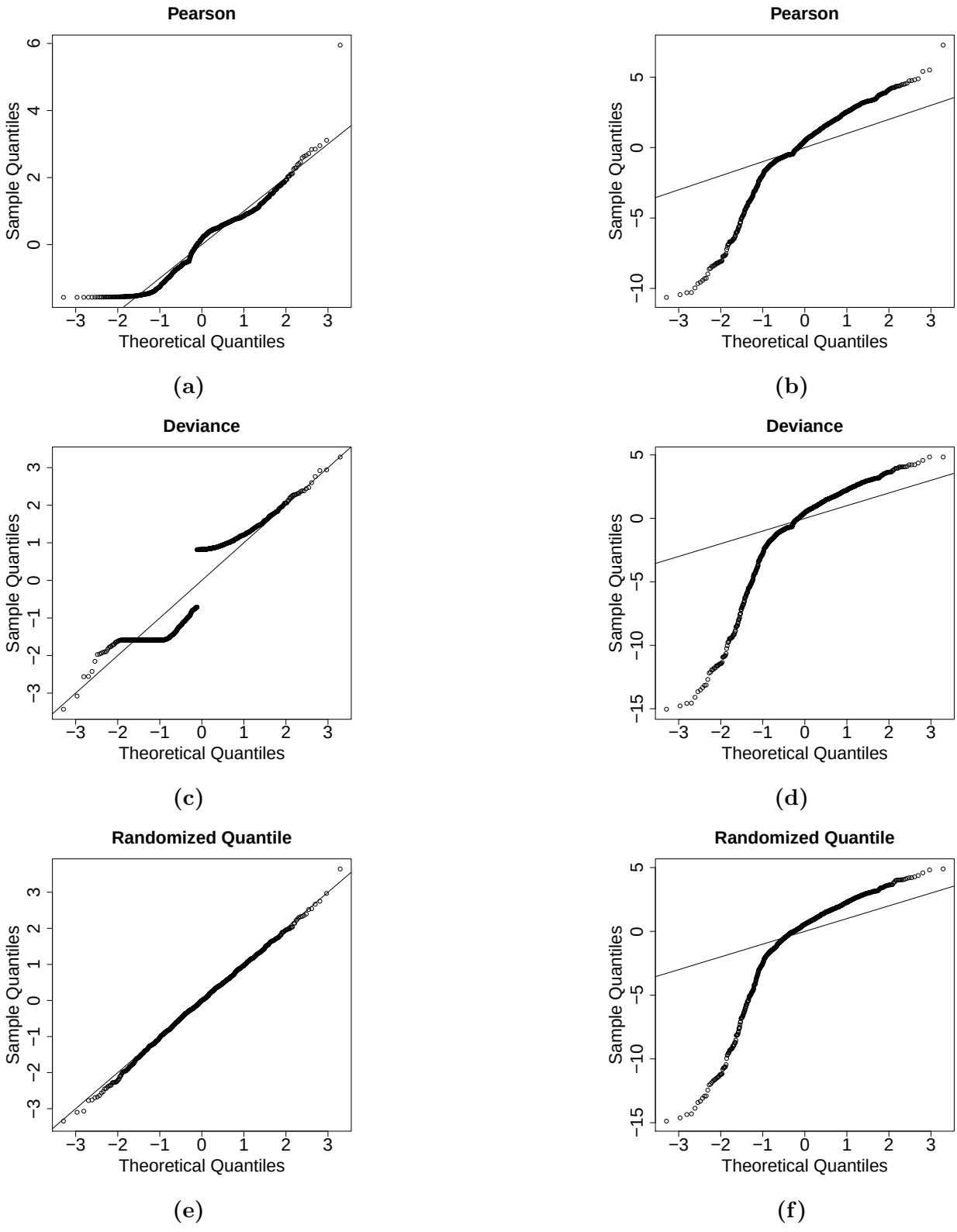


Figure 3.14: Pearson, deviance, and randomized quantile residuals for two models; left panel: $y|x \sim ZIP(\exp(\beta_0+\beta_1 x))$ (true model) and right panel: $y|x \sim Poisson(\exp(\beta_0+\beta_1 x))$ (wrong model)

for randomized quantile residuals are not uniformly distributed, the model can not be chosen as true model. Figure 3.15e shows that that the p-value is almost uniformly distributed for ZIP model and it cane be chosen as true model. The results from other normality tests are almost the same and will be provided upon request. In this case, the optimistic bias also exits.

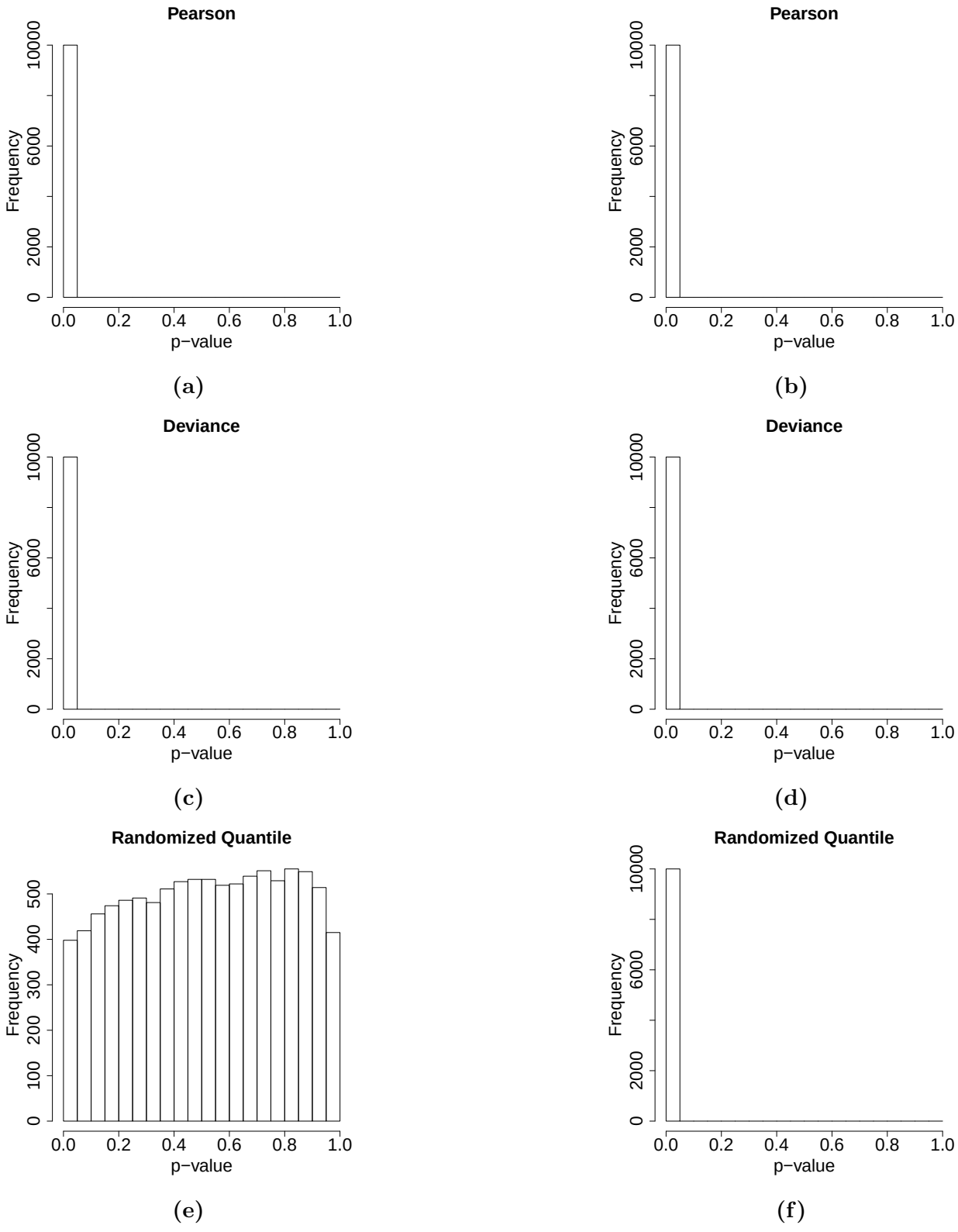


Figure 3.15: P-value from the Wilk-Shapiro test for Pearson, deviance, and randomized quantile residuals for two models; left panel: $y|x \sim ZIP(\exp(\beta_0 + \beta_1 x))$ (true model) and right panel: $y|x \sim Poisson(\exp(\beta_0 + \beta_1 x))$ (wrong model)

Chapter 4

Application to $PM_{2.5}$ Data

In this Chapter, we will compare randomized quantile residuals with the traditional residuals in a real dataset. At first, we give a short background about the study, then we describe the dataset and the variables and finally, we apply the randomized quantile residuals to the dataset.

4.1 Introduction

In multiple recent toxicological and epidemiological studies, particulate matter which are 2.5 μm ($PM_{2.5}$) has been shown to correlate with harmful health outcome [5, 13, 16, 19, 21, 33, 36, 39]. They are mainly generated from combustion processes. Due to the small size of $PM_{2.5}$, they can easily penetrate deeper into lungs and blood streams unfiltered [9], leading to respiratory [10, 34, 36] and cardiovascular diseases [13, 28, 33]. Few studies have investigated the relationship between daily $PM_{2.5}$ exposure and influenza-like-illnesses (ILI). Such studies are necessary to provide evidence for the different effects of $PM_{2.5}$ exposure's influence on the development of influenza within high risk populations (eg. Children, elderly, chronic disease patients, etc.). This Chapter is based on [16] in which influenza-like-illnesses have been shown to have a temporal relationship with $PM_{2.5}$ exposure in Beijing, the capital city of China. That study provides evidence supporting the role of $PM_{2.5}$ exposure in developing influenza-like-illnesses in Beijing, (controlling for the effects of weather conditions, status of the day being a weekend/holiday, month, year) using a generalized additive model (GAM) to flexibly model the nonlinear relationship between the continuous covariates and the outcome variable in the year of 2013. Delayed effect of $PM_{2.5}$ was also considered.

4.2 Data Sources and Descriptions

Influenza-like-illness is defined as daily number of patients who sought medical attention with body temperature more than 38° Celsius with cough or sore throat. The influenza data from Janurary 1, 2013 to December 21, 2013 was retrieved from Beijing Centre of Disease Control (CDC) surveillance system [50]. The data came from 150 level two and three hospitals in Beijing of the national, city, and district level. The hospitals cover all 16 districts in Beijing with data from all outpatients with respiratory disease treatment. The data were reported to Beijing CDC everyday from each hospital using an online system and validated by staffs in the district CDC.

The average daily $PM_{2.5}$ measurements during the study period was retrieved from an air quality monitoring site at the US Embassy in Beijing, located in the Chaoyang district. Though the measurements came from only one location, it measured $PM_{2.5}$ over a long period of time and it was used and validated by other studies as well; see for example [25].

The temperature and relative humidity (considered as confounders of association between $PM_{2.5}$ exposure and influenza-like-illnesses) were obtained through the China Weather Networks out-door weather reports. The daily counts of influenza-like-illnesses, $PM_{2.5}$ levels, and weather data were linked by date [16].

4.3 Data Analysis

Let μ_t be the expected mean number of individuals with ILI on day t and n_t be the population size on day t, estimated by fitting a sigmoid function to the annual population size of Beijing [6, 16, 18, 27]. Suppose the ILI incidence rate be the ratio of μ_t relative to n_t. Let $PM_{2.5,t-p}$ be the the lag of $PM_{2.5}$ by p days, which is the measurement of $PM_{2.5}$ p days ($p = 1, \cdots, 5$) before ILI case report date t. Following the context in generalized additive models (GAMs), let $f_j(x)$, $j = 1, \cdots, 5$ denote the penalized smoothing spline functions for $PM_{2.5}$ at flu season (October-April), non- flu season (May-September), temperature, humidity and month, respectively. Based on the model proposed by Feng et al. [16], we consider the following model

for the dataset

$$\log(\mu_t) = \alpha_0 + \log(n_t) + f_1(PM_{2.5,t-p})I(fluseason_t) + f_2(PM_{2.5,t-p})I(nonfluseason_t) + f_3(temperature_t) + f_4(humidity_t) + f_5(month_t) + \gamma I(weekday_t) \quad (4.1)$$

where α_0 is the intercept and γ is the regression coefficient for weekday. $I(A)$ is the indicator function with $I(A) = 1$ if and only if A is true. Lag of p days is denoted by lag p and is used to investigate the delayed impact of $PM_{2.5}$ on ILI risk. The accumulated exposures of $PM_{2.5}$ on ILI incidence is also explored by averaging over the current day and the previous day, denoted by lag 01 and up to five days (lag 05) before the ILI measurements taken. Four distributions are considered for the response variable: Poisson, negative binomial, Gamma, and inverse Gaussian with log link function. The detailed model Akaikes Information Criterion (AIC) [2] is presented in Table 4.1.

Table 4.1: AIC scores for the Poisson, negative binomial, Gamma, and inverse Gaussian GAMs with log link function based on model in (4.1). The bolded number in the table indicates the model with the smallest AIC.

lag	Poisson	Negative Binomial	Gamma	Inverse Gaussian
lag0	12913.16	4892.359	4887.074	**4844.341**
lag1	12833.69	4897.604	4889.308	4846.355
lag2	12938.79	4900.928	4893.896	4850.798
lag3	13197.37	4901.484	4895.458	4851.314
lag4	13272.44	4901.034	4893.929	4849.434
lag5	13089.75	4900.378	4892.172	4850.023
lag01	12751.81	4892.554	4884.817	4846.206
lag02	12301.98	4883.293	4874.695	4874.695
lag03	12706.88	4895.858	4888.591	4848.433
lag04	12624.70	4896.806	4889.061	4848.628
lag05	12922.24	4897.522	4889.469	4849.264

4.3.1 Negative Binomial Model

First, we compare the Poisson and negative binomial regression, two models that are primarily designed for count data, due to the fact that they provide easier interpretation and more meaningful model for count data. Based on the AICs in Table 4.1, the lag 0 negative binomial GAM has the smaller AIC among all negative binomial and Poisson models. In the lag 0 negative binomial GAM, the variable holiday weekend was not significant, so is removed from further analysis. Before accepting it as the true model, we need to check its residuals. All three types of residuals plotted versus covariates and QQ-plots are presented in Figures 4.1, 4.2, and 4.3. Pearson and deviance are not forming parallel lines as expected, because the response variable has 364 observation with 302 distinct values ranging from 937 to 4603. None of the residuals are normal (p-value < 0.001). The QQ-plots show somewhat slight departure from a negative binomial distribution. The reason is because the response variable is ranging from 937 to 4603, while negative binomial requires having relatively small data points. Hence, negative binomial model does not fit the data very well.

4.3.2 Inverse Gaussian Model

Due to the inadequacy of the Poisson and the negative binomial models for modeling this data, we examine the inverse Gaussian and Gamma models. Based on the AIC presented in Table 4.1, the lag 0 inverse Gaussian GAM offers the smallest AIC. Our next goal is to do a thorough model adequacy checking for lag 0 inverse Gaussian model. Because the variable temperature was not significant, it was removed from any subsequent analysis. The three types of residuals are plotted against different covariates in the model and are presented in Figure 4.4, 4.5, 4.6. Here, because the estimtation of dispersion parameter is very small ($9.3 * 10^{-6}$), we used scaled deviance residuals instead to be able to compare them with other types of residuals. Pearson residuals are not normally distributed (p-value $< .001$). However, from Figures 4.5f and 4.6, deviance (p-value $= .24$) and randomized quantile residuals (p-value$= .26$) are normally distributed, which suggests the adequacy of the inverse Gaussian model.

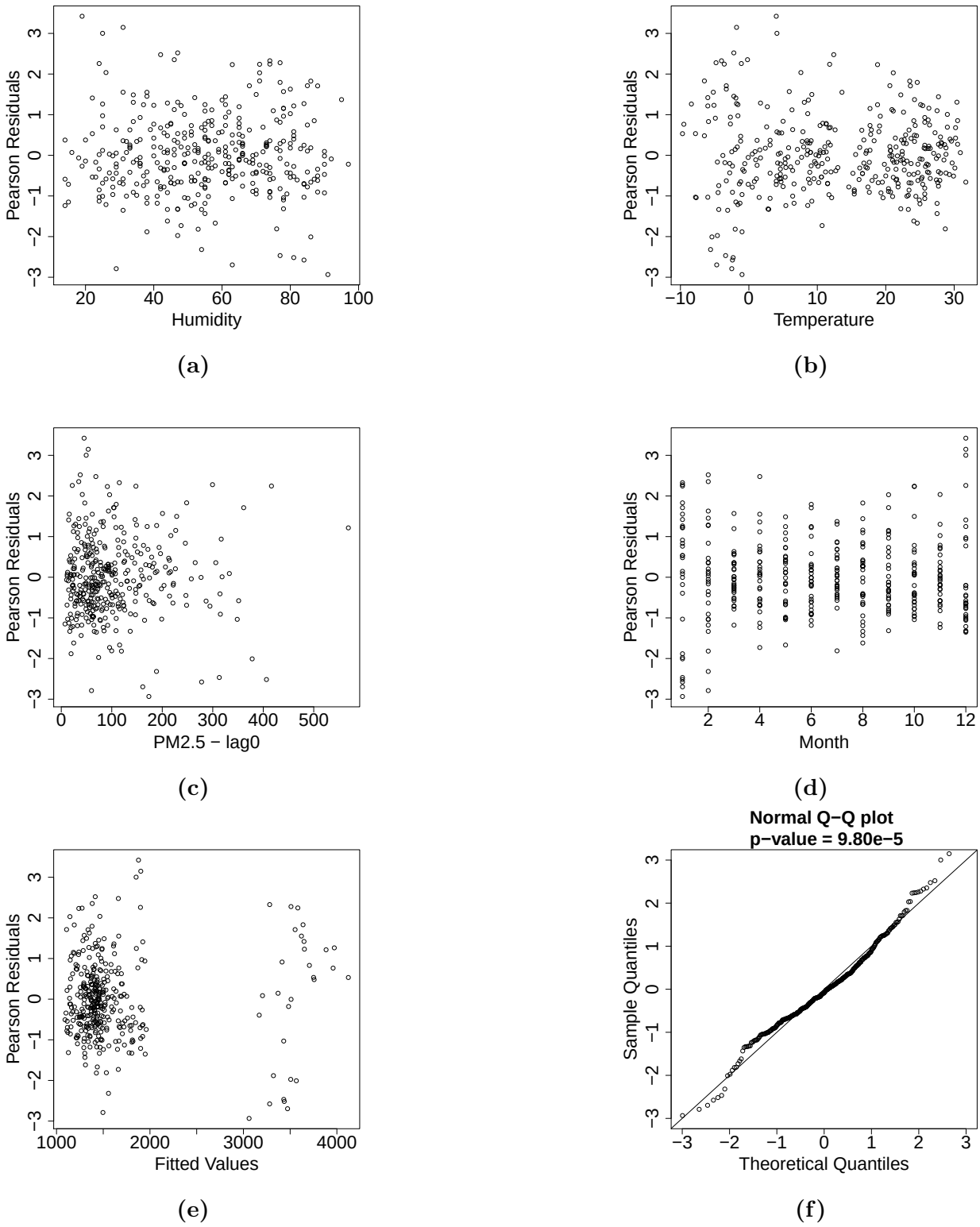


Figure 4.1: Pearson residuals versus each significant covariate in the lag 0 negative binomial model and their QQ-plot.

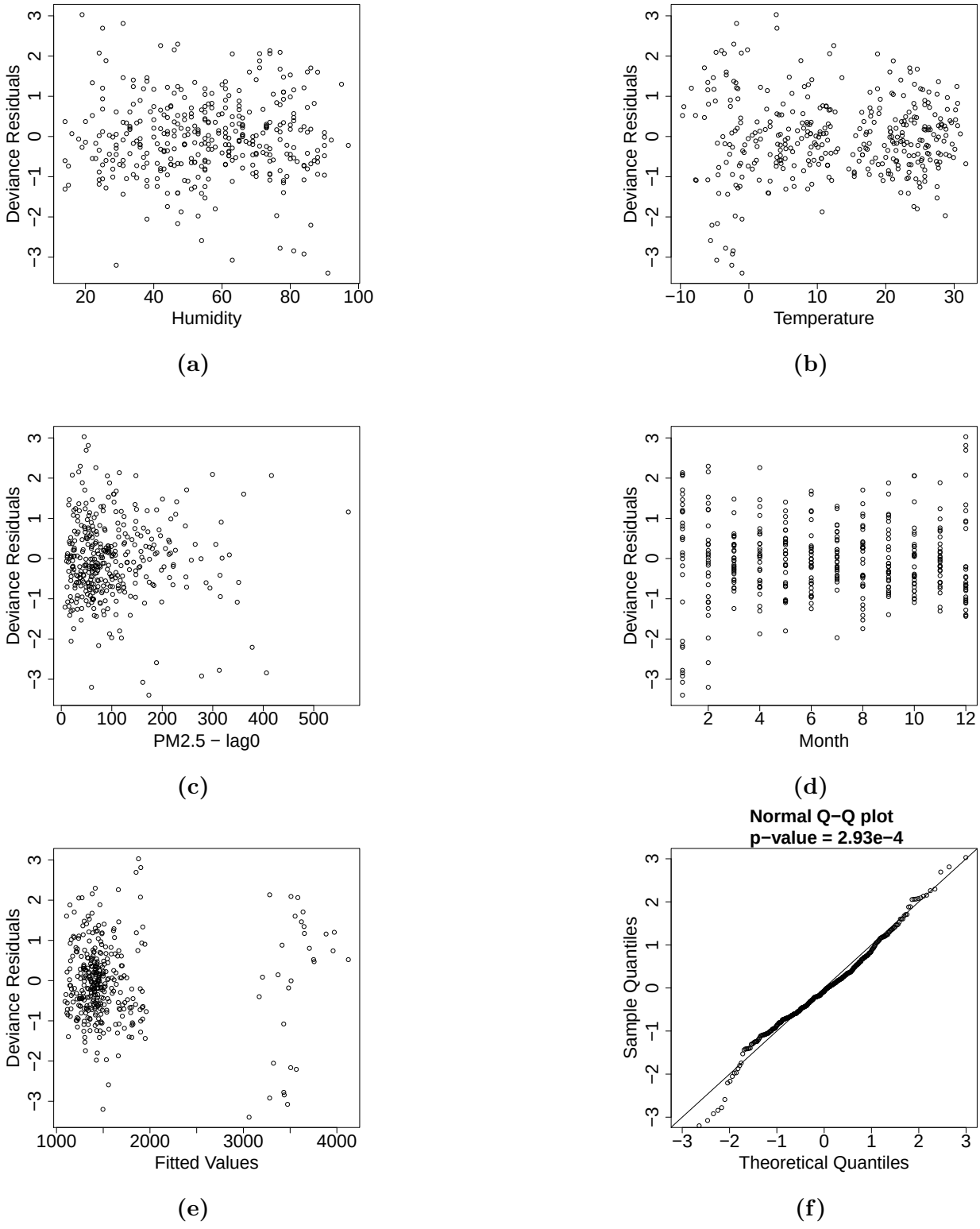


Figure 4.2: Deviance residuals versus each significant covariate in the lag 0 negative binomial model and their QQ-plot

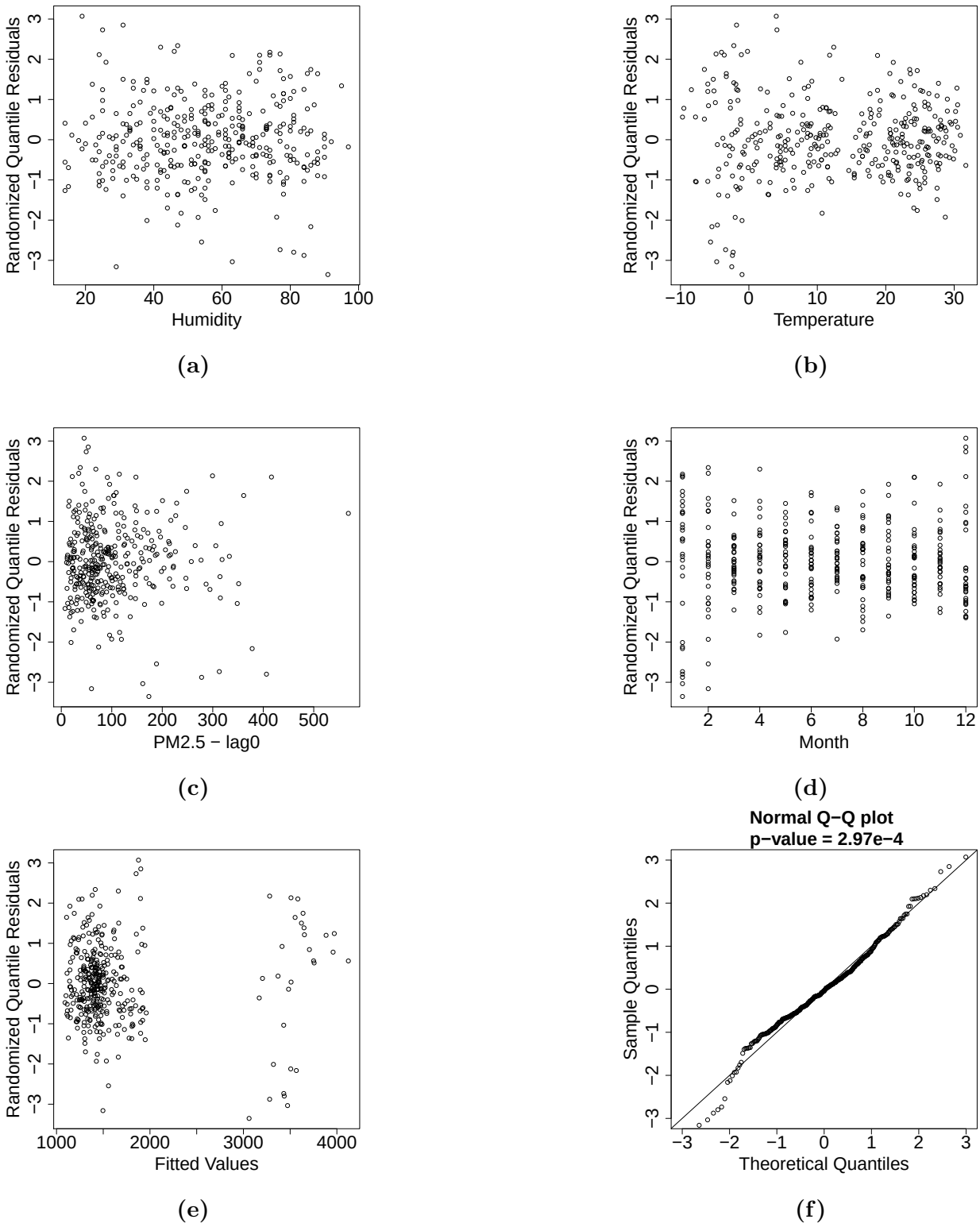
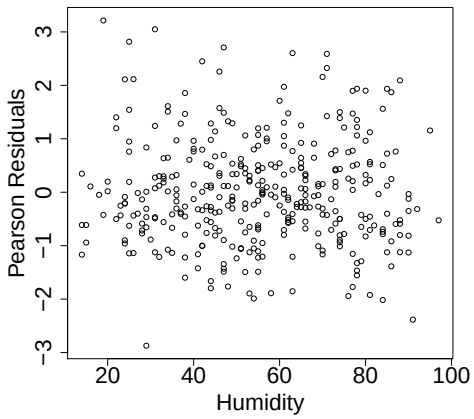
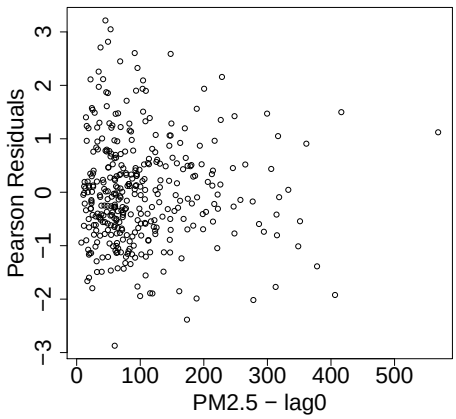


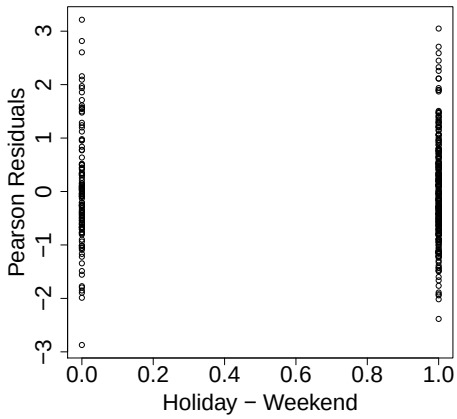
Figure 4.3: Randomized Quantile residuals versus each significant covariate in the lag 0 negative binomial model and their QQ-plot.



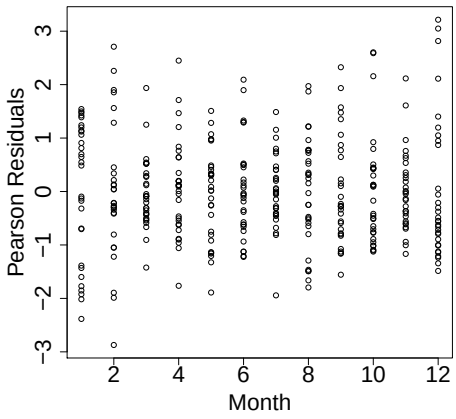
(a)



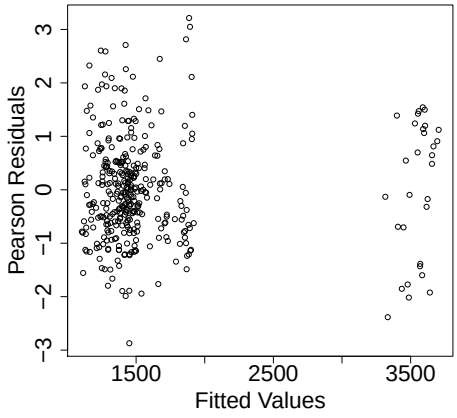
(b)



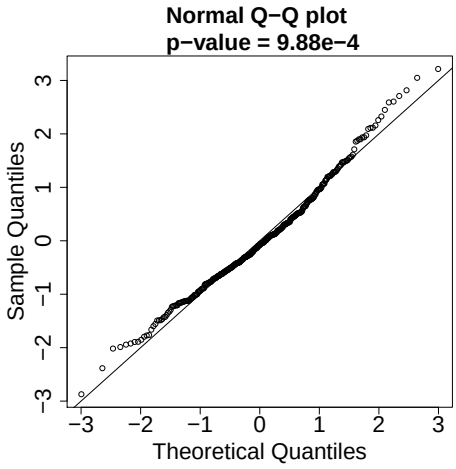
(c)



(d)

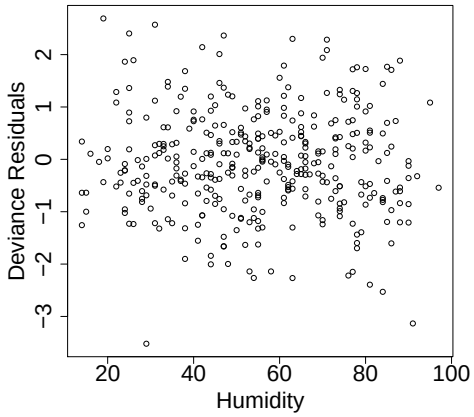


(e)

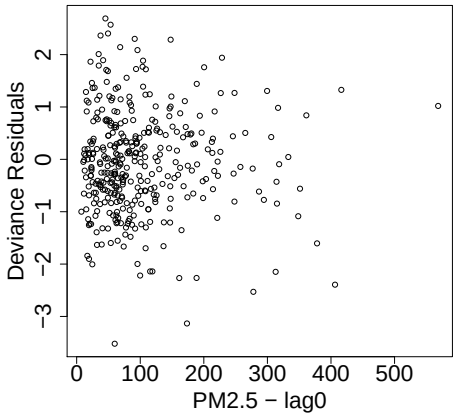


(f)

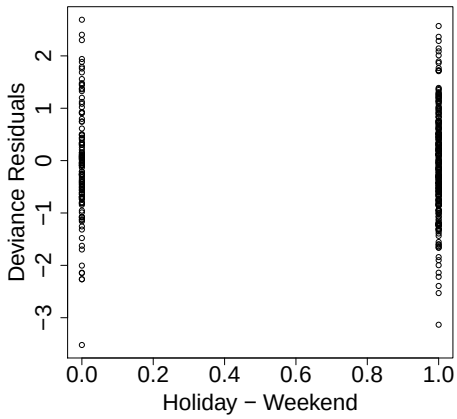
Figure 4.4: Pearson residuals versus each significant covariate in the lag 0 inverse Gaussian model and their QQ-plot



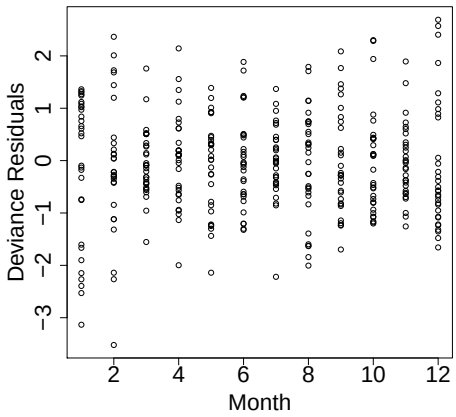
(a)



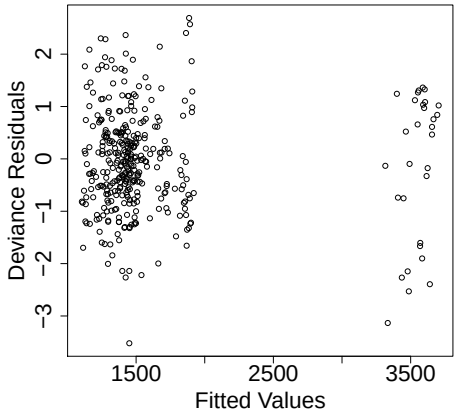
(b)



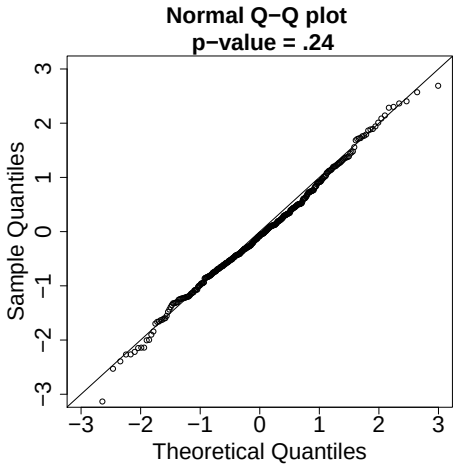
(c)



(d)

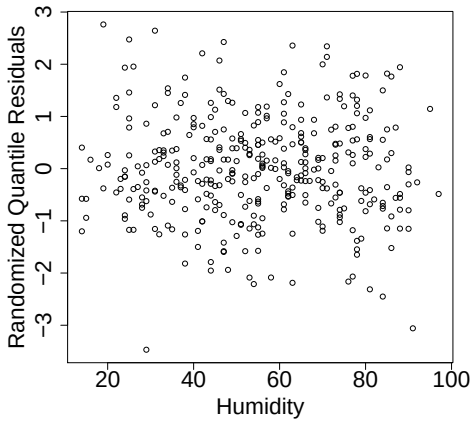


(e)

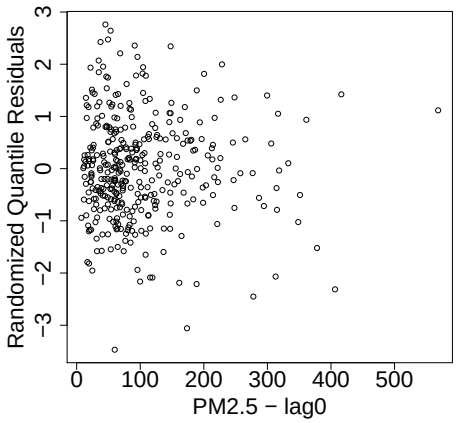


(f)

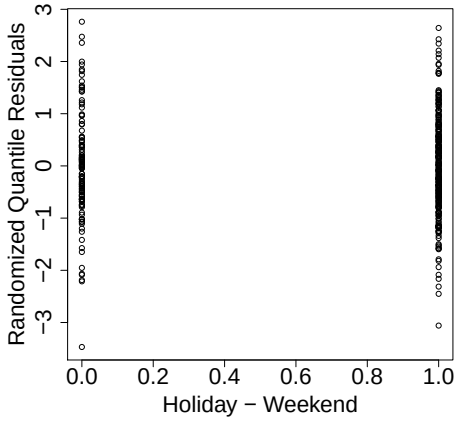
Figure 4.5: Deviance residuals versus each significant covariate in the lag 0 inverse Gaussian model and their QQ-plot



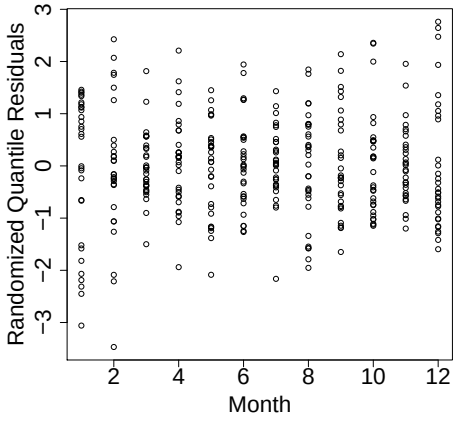
(a)



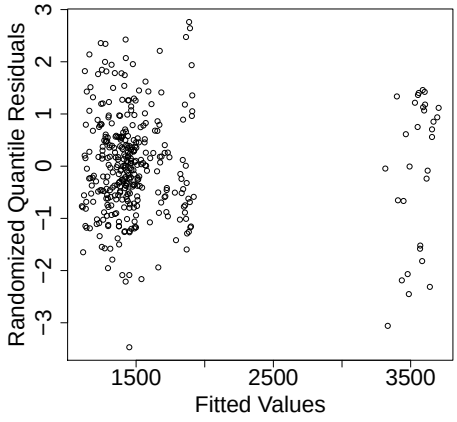
(b)



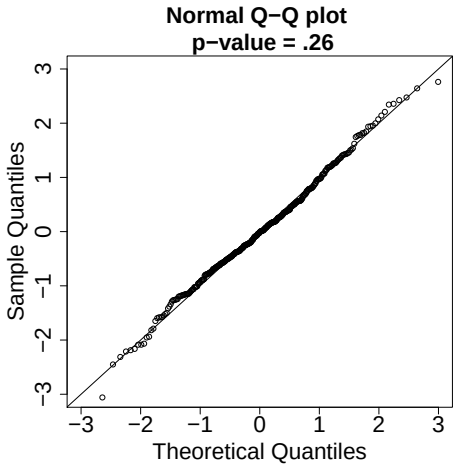
(c)



(d)



(e)



(f)

Figure 4.6: Randomized Quantile residuals versus each significant covariate in the lag 0 inverse Gaussian model and their QQ-plot

Chapter 5

Conclusion and Future Work

In this thesis, we reviewed randomized quantile residuals and theoretically justified the normality of this type of residual and compared its performance with the traditional ones, Pearson and deviance residuals, through a set of simulation studies. This thesis reinforces that the traditional residuals are not well calibrated and fail to assist in model diagnosis, especially for modeling discrete outcome variables. In count data, the residual plots are typically not normally distributed given the true model, which makes it difficult to visually inspect the model fit. On the other hand, randomized quantile residuals are well calibrated and can be used for a wide range of distributions, which provides a unified way for model diagnosis. We theoretically proved that randomized quantile residuals are exactly standard normal, aside from the variability in the estimation of the parameters. This provides a unified way of simply plotting the randomized quantile residuals against predicted values or the covariates as well as their QQ-plots for visually checking the model adequacy. Another significant advantage of randomized quantile residuals over the traditional ones is their simple and unified definition, which only requires knowing the cumulative distribution function of the response variable.

Our simulation study demonstrated the excellent performance of randomized quantile residuals. In particular, the randomized quantile residuals can detect non-linearity in GLM, such as Poisson, negative binomial, and Gamma distribution, whereas Pearson and deviance residuals are unable to give us confidence in the model diagnosis when the model is true. Aside from the residual plots versus covariates or predicted values that can discover non-linearity in the model, their QQ-plots are also able to highlight the non-linearity in the covariate effect, while deviance and Pearson residuals fail to confirm the true model in most of the cases. Our simulation studies showed that randomized quantile residuals are able

to identify overdispersion when the true model is negative binomial but fitting a Poisson model. The simulations also demonstrated that randomized quantile residuals can recognize zero-inflation in the response variable.

In practice, researchers often use Pearson and deviance χ^2-test to assess overall goodness-of-fit. Therefore, we also examined if the χ^2-test can be applied to randomized quantile residuals and how it performs as compared to Pearson and deviance. Some of the results of applying χ^2 test to the three kinds of residuals are shown in Appendix A. Our simulation studies indicate that in most of the cases, the p-values of the χ^2-tests are not uniformly disributed. Therefore, we do not suggest applying χ^2-test to any of the three residuals (Pearson, deviance, and randomized quantile residuals) to check the overall fit of models. As an alternative way to χ^2-test, we recommend people to use Wilk-Shapiro test or other types of normality tests to the residuals as a GOF test. Simulation showed that Wilk-Shapiro normality test works well and can be used as an alternative way to χ^2-test.

Despite all the above-mentioned improvements of randomized quantile residuals over traditional ones, they depend on the choice of the uniform random variable. Dunn and Smyth [15] suggested calculating four realizations of randomized quantile residuals and then disregard any inconsistent pattern among them.

In simulations, we observed that sometimes the actual observations appear to be more predictable by the model, known as optimistic bias. Specifically, in negative binomial and zero-inflated Poisson regression, for which one needs to estimate dispersion for the former and probability of zeros in latter, the p-value for Wilk-Shapiro normality test is not uniformly distributed having more chances to accept the true model. Although this problem is not serious when the mean is large relative to the dispersion or zero-inflation parameter, for other cases, optimistic bias should be properly addressed. This problem may be more severe in more complicated models such as latent variable models for modeling dependent data. We speculate that cross-validation methods such as leave-one-out cross-validation (LOOCV) might alleviate this problem. We also believe that employing Bayesian methods such as importance sampling procedure can be used to overcome the optimistic biased (see [24, 35] for example). There are other types of residuals such as studentized and Anscombe residuals that have been used in literature. The comparison between these residuals and randomized

quantile residual can be a good topic for future research as well.

Bibliography

[1] Agresti, A. (2015), *Foundations of linear and generalized linear models*, John Wiley & Sons.

[2] Akaike, H. (1992), "Information theory and an extension of the maximum likelihood principle," in *Breakthroughs in statistics*, Springer, pp. 610–624.

[3] Anderson, T. W. and Darling, D. A. (1952), "Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes," *The annals of mathematical statistics*, 193–212.

[4] — (1954), "A test of goodness of fit," *Journal of the American statistical association*, 49, 765–769.

[5] Atkinson, R. W., Fuller, G. W., Anderson, H. R., Harrison, R. M., and Armstrong, B. (2010), "Urban ambient particle metrics and health: a time-series analysis," *Epidemiology*, 21, 501–511.

[6] Banks, R. B. (2013), *Growth and diffusion phenomena: Mathematical frameworks and applications*, vol. 14, Springer Science & Business Media.

[7] Benjamin, M. A., Rigby, R. A., and Stasinopoulos, M. D. (2003), "Generalized autoregressive moving average models," *Journal of the American Statistical association*, 98, 214–223.

[8] Cameron, C. A. and Trivedi, P. K. (2013), *Regression analysis of count data*, vol. 53, Cambridge university press.

[9] Chan, T. L. and Lippmann, M. (1980), "Experimental measurements and empirical modelling of the regional deposition of inhaled particles in humans," *The American Industrial Hygiene Association Journal*, 41, 399–409.

[10] Chauhan, A. J. and Johnston, S. L. (2003), "Air pollution and infection in respiratory illness," *British medical bulletin*, 68, 95–112.

[11] Cox, D. R. and Snell, E. J. (1968), "A general definition of residuals," *Journal of the Royal Statistical Society. Series B (Methodological)*, 248–275.

[12] Dallal, G. E. and Wilkinson, L. (1986), "An analytic approximation to the distribution of Lilliefors's test statistic for normality," *The American Statistician*, 40, 294–296.

[13] Delfino, R. J., Sioutas, C., and Malik, S. (2005), "Potential role of ultrafine particles in associations between airborne particle mass and cardiovascular health," *Environmental health perspectives*, 934–946.

[14] Dunn, P. K. (2004), "Occurrence and quantity of precipitation can be modelled simultaneously," *International Journal of Climatology*, 24, 1231–1239.

[15] Dunn, P. K. and Smyth, G. K. (1996), "Randomized quantile residuals," *Journal of Computational and Graphical Statistics*, 5, 236–244.

[16] Feng, C., Li, J., Sun, W., Zhang, Y., and Wang, Q. (2016), "Impact of ambient fine particulate matter (PM 2.5) exposure on the risk of influenza-like-illness: a time-series analysis in Beijing, China," *Environmental Health*, 15, 1.

[17] Gross, J. and Ligges, U. (2015), *nortest: Tests for Normality*, r package version 1.0-4.

[18] Grübler, A. (1991), "Diffusion: long-term patterns and discontinuities," in *Diffusion of Technologies and Social Behavior*, Springer, pp. 451–482.

[19] Janssen, N., Fischer, P., Marra, M., Ameling, C., and Cassee, F. (2013), "Short-term effects of PM 2.5, PM 10 and PM 2.5–10 on daily mortality in the Netherlands," *Science of the Total Environment*, 463, 20–26.

[20] Jorgensen, B. (1987), "Exponential dispersion models," *Journal of the Royal Statistical Society. Series B (Methodological)*, 127–162.

[21] Laden, F., Neas, L. M., Dockery, D. W., and Schwartz, J. (2000), "Association of fine particulate matter from different sources with daily mortality in six US cities." *Environmental health perspectives*, 108, 941.

[22] Lambert, D. (1992), "Zero-inflated Poisson regression, with an application to defects in manufacturing," *Technometrics*, 34, 1–14.

[23] Lee, A. H., Wang, K., and Yau, K. K. (2001), "Analysis of zero-inflated Poisson data incorporating extent of exposure," *Biometrical Journal*, 43, 963.

[24] Li, L., Qiu, S., Zhang, B., and Feng, C. X. (2015), "Approximating cross-validatory predictive evaluation in Bayesian latent variable models with integrated IS and WAIC," *Statistics and Computing*, 1–17.

[25] Liang, Y., Fang, L., Pan, H., Zhang, K., Kan, H., Brook, J. R., and Sun, Q. (2014), "PM 2.5 in Beijing–temporal pattern and its association with influenza," *Environmental Health*, 13, 1.

[26] Lilliefors, H. W. (1967), "On the Kolmogorov-Smirnov test for normality with mean and variance unknown," *Journal of the American Statistical Association*, 62, 399–402.

[27] Marchetti, C., Meyer, P. S., and Ausubel, J. H. (1996), "Human population dynamics revisited with the logistic model: how much can be modeled and predicted?" *Technological Forecasting and Social Change*, 52, 1–30.

[28] Maté, T., Guaita, R., Pichiule, M., Linares, C., and Díaz, J. (2010), "Short-term effect of fine particulate matter (PM 2.5) on daily mortality due to diseases of the circulatory system in Madrid (Spain)," *Science of the Total Environment*, 408, 5750–5757.

[29] McCullagh, P. and Nelder, J. A. (1989), *Generalized linear models*, vol. 37, CRC press.

[30] Nelder, J. and Wedderburn, R. (1972), "Generalized linear models," .

[31] Ospina, R. and Ferrari, S. L. (2012), "A general class of zero-or-one inflated beta regression models," *Computational Statistics & Data Analysis*, 56, 1609–1623.

[32] Pierce, D. A. and Schafer, D. W. (1986), "Residuals in generalized linear models," *Journal of the American Statistical Association*, 81, 977–986.

[33] Polichetti, G., Cocco, S., Spinali, A., Trimarco, V., and Nunziata, A. (2009), "Effects of particulate matter (PM 10, PM 2.5 and PM 1) on the cardiovascular system," *Toxicology*, 261, 1–8.

[34] Qiu, H., Yu, I. T.-s., Tian, L., Wang, X., Tse, L. A., Tam, W., and Wong, T. W. (2012), "Effects of coarse particulate matter on emergency hospital admissions for respiratory diseases: a time-series analysis in Hong Kong," *Environmental health perspectives*, 120, 572.

[35] Qiu, S., Feng, C. X., and Li, L. (2016), "Approximating Cross-validatory Predictive P-values with Integrated IS for Disease Mapping Models," *arXiv preprint arXiv:1603.07668*.

[36] Raaschou-Nielsen, O., Andersen, Z. J., Beelen, R., Samoli, E., Stafoggia, M., Weinmayr, G., Hoffmann, B., Fischer, P., Nieuwenhuijsen, M. J., Brunekreef, B., et al. (2013), "Air pollution and lung cancer incidence in 17 European cohorts: prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE)," *The lancet oncology*, 14, 813–822.

[37] Razali, N. M. and Wah, Y. B. (2011), "Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests," *Journal of Statistical Modeling and Analytics*, 2, 21–33.

[38] Rigby, R. A. and Stasinopoulos, M. D. (2005), "Generalized additive models for location, scale and shape," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54, 507–554.

[39] Schwartz, J., Dockery, D. W., and Neas, L. M. (1996), "Is daily mortality associated specifically with fine particles?" *Journal of the Air & Waste Management Association*, 46, 927–939.

[40] Shapiro, S. S. and Francia, R. (1972), "An approximate analysis of variance test for normality," *Journal of the American Statistical Association*, 67, 215–216.

[41] Shapiro, S. S. and Wilk, M. B. (1965), "An analysis of variance test for normality (complete samples)," *Biometrika*, 52, 591–611.

[42] Shapiro, S. S., Wilk, M. B., and Chen, H. J. (1968), "A comparative study of various tests for normality," *Journal of the American Statistical Association*, 63, 1343–1372.

[43] Smyth, G. K. (2003), "Pearson's goodness of fit statistic as a score test statistic," *Lecture Notes-Monograph Series*, 115–126.

[44] Stephens, M. A. (1974), "EDF statistics for goodness of fit and some comparisons," *Journal of the American statistical Association*, 69, 730–737.

[45] — (1986), "Tests based on EDF statistics," *Goodness-of-fit Techniques*, 68, 97–193.

[46] Thode, H. C. (2002), *Testing for normality*, vol. 164 of *Statistics: textbooks and monographs*, CRC press.

[47] Venables, W. N. and Ripley, B. D. (2013), *Modern applied statistics with S-PLUS*, Springer Science & Business Media.

[48] Watson, G. S. (1961), "Goodness-of-fit tests on a circle," *Biometrika*, 48, 109–114.

[49] — (1962), "Goodness-of-fit tests on a circle. II," *Biometrika*, 49, 57–63.

[50] Yang, P., Thompson, M. G., Ma, C., Shi, W., Wu, S., Zhang, D., and Wang, Q. (2014), "Influenza vaccine effectiveness against medically-attended influenza illness during the 2012–2013 season in Beijing, China," *Vaccine*, 32, 5285–5289.

[51] Zeileis, A., Kleiber, C., and Jackman, S. (2008), "Regression Models for Count Data in R," *Journal of Statistical Software*, 27.

Appendix.

χ^2-Tests

Researchers usually prefer to use Pearson and deviance χ^2-test to assess model fit. In this chapter, we will investigate these tests via same simulation studies that we did in Chapter 3. Assume

$$D^*\left(\boldsymbol{y};\hat{\boldsymbol{\mu}}\right)=\frac{D\left(\boldsymbol{y};\hat{\boldsymbol{\mu}}\right)}{\phi}=\frac{\sum_i d_i^2}{\phi} \tag{A.1}$$

$$X^2=\sum_i r_i^2 \tag{A.2}$$

$$Q^2=\sum_i q_i^2 \tag{A.3}$$

Because q_i are normally distributed, Q has a χ^2 distribution, if the model is true. In normal case $D^*\left(\boldsymbol{y};\hat{\boldsymbol{\mu}}\right)$ and X^2 has also χ^2 distribution if true parameters are used. In non-normal case, there are lots of cases for which $D^*\left(\boldsymbol{y};\hat{\boldsymbol{\mu}}\right)$ and X^2 have approximate χ^2 distribution. In those cases, it is often believed that the degrees of freedom is $n-p$, where p is the number of model parameters needed to be estimated in the model, however it may not be true due to the fact that typically estimation of parameters are used [29]. Even though there are cases for which the distribution of $D^*\left(\boldsymbol{y};\hat{\boldsymbol{\mu}}\right)$ and X^2 are unknown, Pearson and deviance χ^2-tests are widely used to assess overall GOF. To see if these tests can differentiate between the true model and the wrong model, we calculate p-value from the χ^2-test for different types of residuals for replicated simulations in Chapter 3 in each scenario and problem.

Poisson

In Poisson case, after simulation 10000 datasets from $y_i \sim Poisson(\exp(x_i^2))$, and fitting two models, linear and quadratic Poisson, we calculate p-value from the three types of χ^2-test. The histogram for these p-values are depicted in Figure A.1. The deviance can not tell the difference between the true model and the wrong model. For wrong model, the p-value for

both Pearson and randomized quantile residuals are very small, indicating correctly that the model does not fit the data very well. For true model, the p-value for both Pearson and randomized quantile residuals are uniformly distributed, confirming that the model is indeed the true model. Hence, both Pearson and randomized quantile χ^2-tests can choose the true model as compared to the wrong model.

Although in Poisson case, Pearson and randomized quantile χ^2-tests could choose the true model in comparison with the wrong model, in other cases, all three χ^2-tests fail to assist in choosing the true model (See Figures A.2 - A.5). Thus, applying χ^2-test to any of the residuals is not advisable. As an alternative way, we recommend using Wilk-Shapiro normality test to assess the model fit.

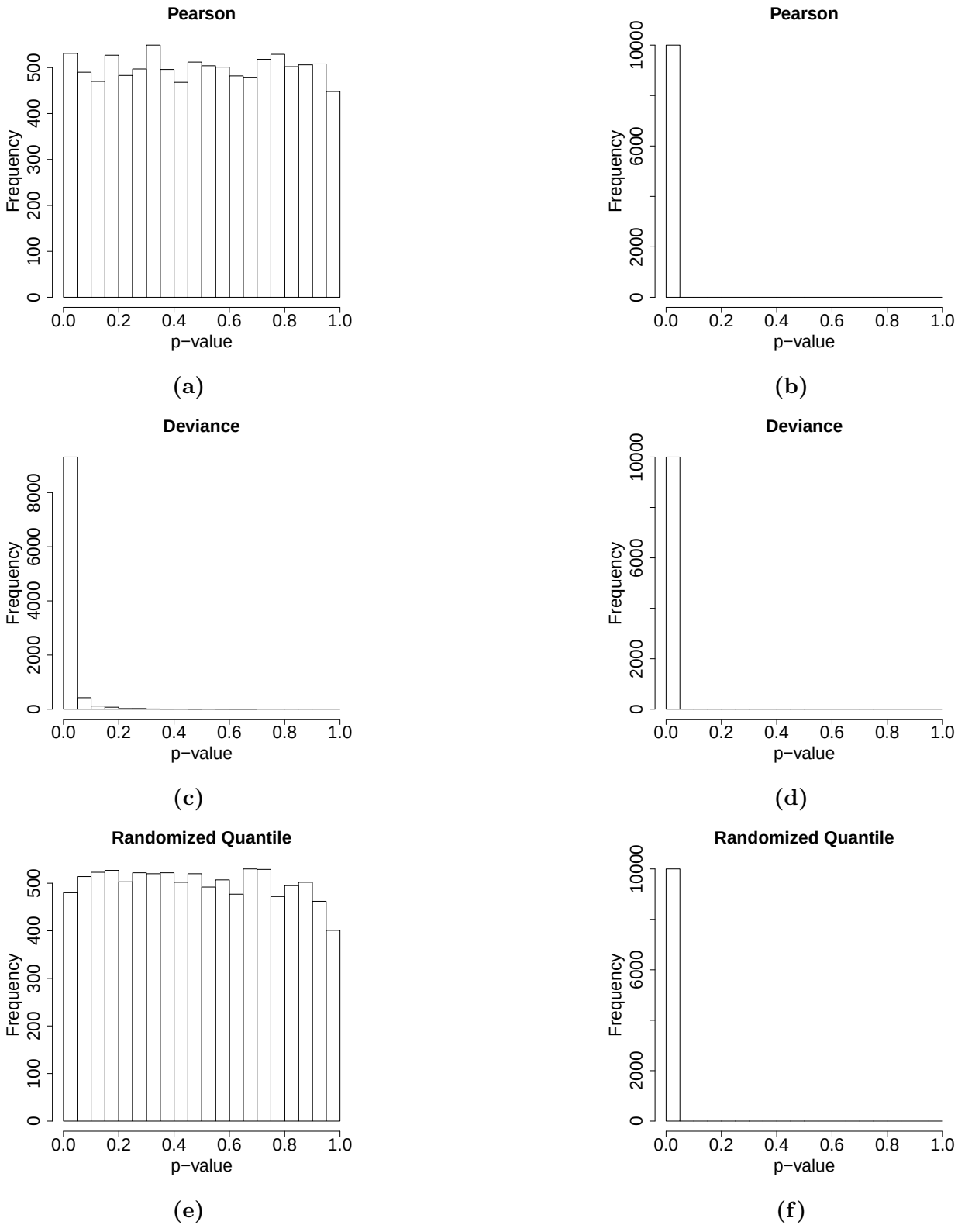


Figure A.1: The p-value from Pearson, deviance, and randomized quantile GOF-tests for two models; left panel: $y|x \sim Poisson(\exp(\beta_0+\beta_1 x^2))$ (true model) and right panel: $y|x \sim Poisson(\exp(\beta_0 + \beta_1 x))$ (wrong model)

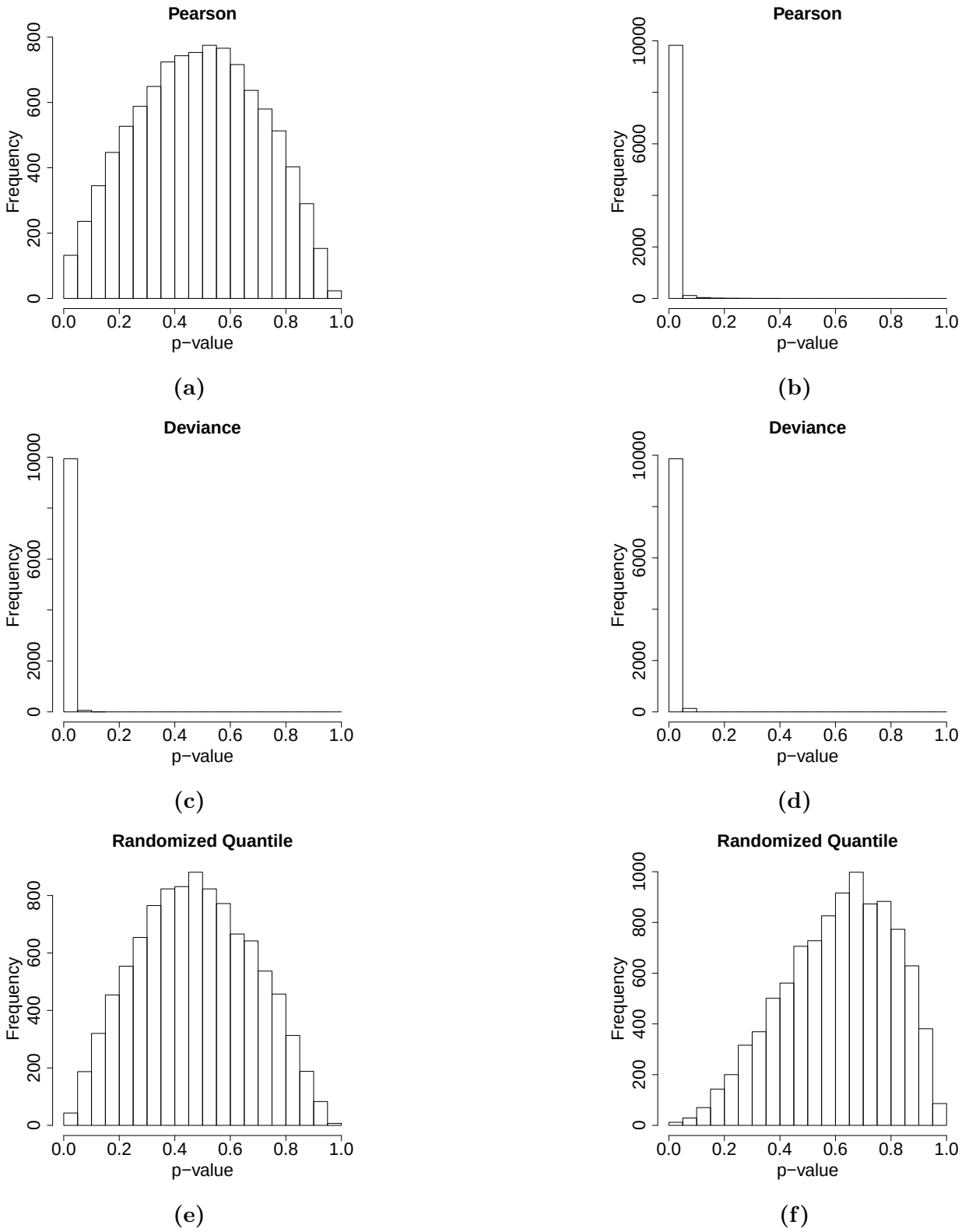


Figure A.2: The p-value from Pearson, deviance, and randomized quantile GOF-tests for two models; left panel: $y|x \sim NB(\exp(\beta_0 + \beta_1 x^2), k)$ (true model) and right panel: $y|x \sim NB(\exp(\beta_0 + \beta_1 x), k)$ (wrong model)

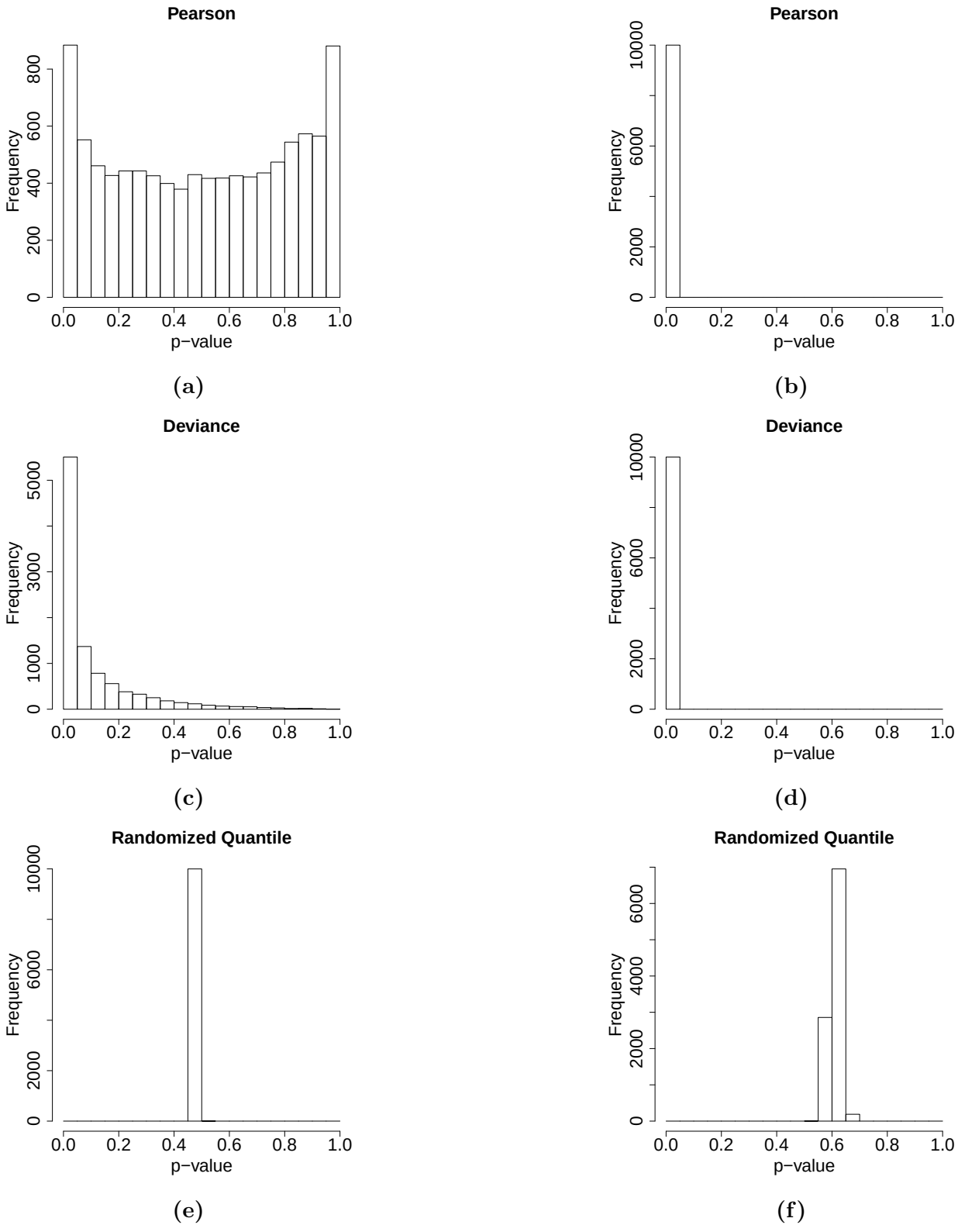


Figure A.3: P-value from Pearson, deviance, and randomized quantile GOF-tests for two models; left panel: $Gamma(k, \exp(\beta_0 + \beta_1 x^2))$ (true model) and right panel: $y|x \sim Gamma(k, \exp(\beta_0 + \beta_1 x))$ (wrong model)

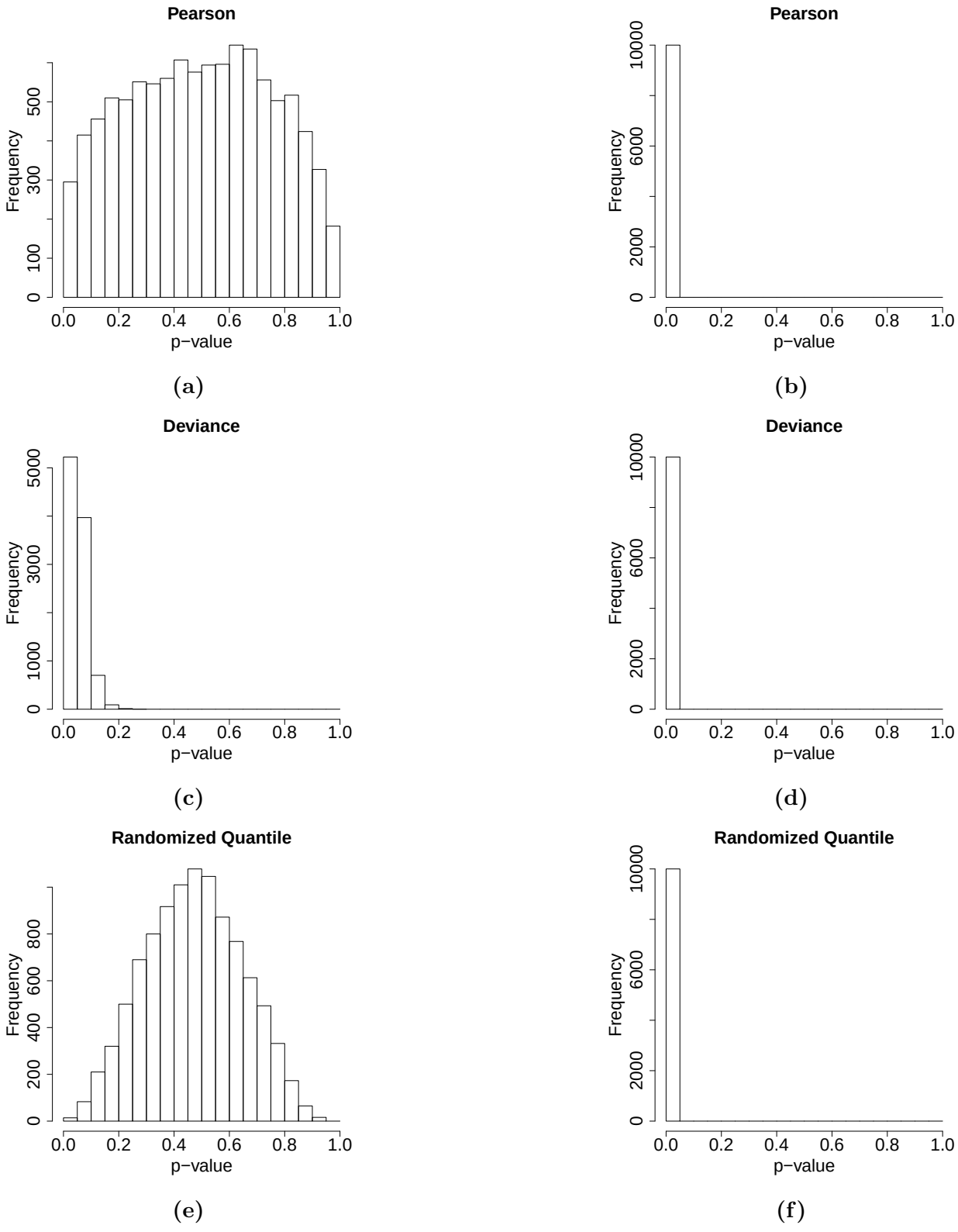


Figure A.4: P-value from Pearson, deviance, and randomized quantile GOF-tests for two models; left panel: $y|x \sim NB(\exp(\beta_0 + \beta_1 x), k)$ (true model) and right panel: $y|x \sim Poisson(\exp(\beta_0 + \beta_1 x))$ (wrong model)

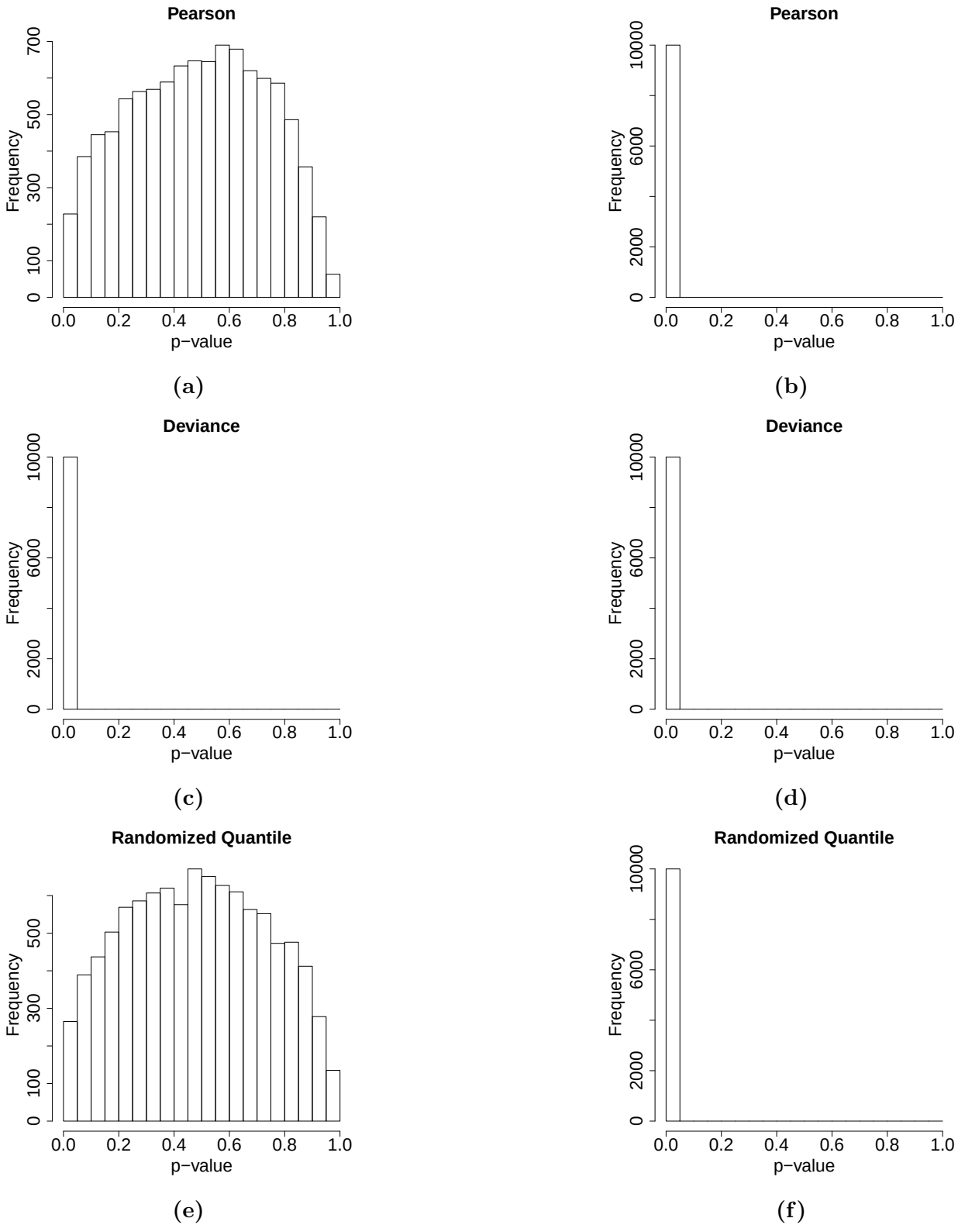


Figure A.5: P-value from Pearson, deviance, and randomized quantile Gof-tests for two models; left panel: $y|x \sim ZIP(\exp(\beta_0 + \beta_1 x))$ (true model) and right panel: $y|x \sim Poisson(\exp(\beta_0 + \beta_1 x))$ (wrong model)